Universidad Nacional Autónoma de México

Facultad de Ciencias

Técnicas de aprendizaje de máquina para la predicción de movimientos del mercado accionario mexicano

T E S I S

Que para obtener el título de
Actuario

Presenta
Julio Antonio Rojas Vallarta.

Director de tesis:
Act. Eduardo Selim Martínez Mayorga.

Ciudad Universitaria, Cd. Mx, Octubre 2021

A mis padres,
a mis amigos.

Agradecimientos

A mis padres Araceli Vallarta y Julio Rojas, por haberme apoyado en cada etapa desde el inicio y hasta la conclusión de mi proyecto.

A mis tíos, en especial a Jazmín Vallarta, por siempre estar ahí brindando algún consejo.

A mis primos, en particular a Ivette Vallarta y Daniel Vallarta, por siempre compartir una sonrisa o una buena charla.

A mis hermanos por elección Eduardo Onofre y Sebastían Tamés por estar ahí siempre dispuestos a platicar, jugar y desvelarse.

A mis profesores, que no fueron pocos y quizá un poco más de los debidos, pero en especial al M. en F. Alberto Cadena, Mae. Francisco Carmona, Act. Francisco Ramírez, Act. Patricia Luna, M. en C. Elena Oteyza, M. en C. Fernanda Gil Leyva, M. en I. Karen Obeid, Dr. Miguel Angel Corona, Mat. Margarita Chávez, quienes de una u de otra forma me ayudaron a formarme en términos académicos y profesionales.

A mi estimado Act. Eduardo Selim Martínez Mayorga con respeto, admiración y agradecimiento por su apoyo, consejo y dirección en este trabajo.

A mis sinodales, Dra. Lizbeth Naranjo Albarrán, Dr. Gonzalo Pérez de la Cruz, M. en C. Claudia Ivonne Juarez Gallegos y M. en C. Jaime Vázquez Alamilla, por sus observaciones y recomendaciones para el enriquecimiento de este trabajo.

A mis amigos, Gabriela Morales, Carlos Torres, Yutsil Flores, Juan Carlos García, Javier Canales, Roger López y por supuesto a Luis Zárate, que sin ellos mi etapa universitaria hubiese sido completamente diferente.

Al "crew" IMEFu, Mauricio Rosales, Karla Gómez de León, Fernando Arceo, Andrés Goca, Andrea García, Erick Mora, Ariadna Ríos y Rafael Acuña, por el aprendizaje y confianza brindados.

Julio Antonio Rojas Vallarta

Índice general

Introducción

El hombre desde sus orígenes ha estado interesado en poder predecir o conocer el futuro, se han hecho diversas prácticas, una de las más famosas era la consulta al oráculo con los antiguos griegos. No es de extrañarse que aún ahora se quiera poder hacer esto y más aún en el ramo financiero, donde siempre se mantiene un constante movimiento de miles de millones de dólares.

Es por ello que para los inversionistas, grandes o pequeños, realizar la predicción del precio futuro de, en este caso, índices accionarios y sus movimientos; sea una de las actividades de mayor interés y una de las que presenta más retos desde la perspectiva de predicción de series de tiempo. Contar con estrategias precisas de los movimientos de índices accionarios es importante para desarrollar estrategias de trading efectivas; con estas predicciones los inversionistas se pueden cubrir contra potenciales riesgos de mercado y los especuladores tener la oportunidad de obtener alguna ganancia comprando o vendiendo el índice accionario.

Adicionalmente, predecir dentro del mercado accionario se considera una de las actividades más desafiantes y estresantes en el proceso de predicción de series de tiempo financieras, ya que el mercado de acciones es esencialmente dinámico, no-lineal, intrincado, no-paramétrico y cuasi-caótico por naturaleza. Dicho mercado está afectado por muchos factores macroeconómicos tales como eventos socio-políticos, condiciones económicas generales, expectativas de los inversionistas, movimientos de otros mercados de acciones, etc.

Por supuesto, al ser México una economía emergente, es de interés estudiar una medida de predictibilidad de los movimientos de los índices accionarios tales como el Índice de Precios y Cotizaciones (IPC), así como el Índice México (INMEX), sin embargo, también pudiendo comparar el desempeño de estos dos índices contra uno de una economía mucho más sólida, como es el índice Standard & Poor's 500, el cual refleja en buena medida el desempeño de la economía estadounidense.

Los índices IPC e INMEX ha tenido la característica de tener rendimientos con alta volatilidad. Dicha volatilidad es atractiva para algunos inversionistas con esperanza de altos rendimientos, pero como métricas de la viabilidad de la economía mexicana no necesaria-

mente es una buena señal.

Por otro lado, en 1959 se acuñó por primera vez el término "aprendizaje automático" o "aprendizaje máquina" por parte de Arthur Samuel, pionero estadounidense en los juegos informáticos y la inteligencia artificial, mientras trabajaba para IBM. El aprendizaje máquina surgió de la búsqueda de la inteligencia artificial, y aunque no se ha conseguido aún una entidad autónoma que pueda llamarse inteligencia artificial per-se, se han hecho varios descubrimientos que pueden aplicarse a más de un solo campo, desde el entretenimiento, la genética, la seguridad informática, y por supuesto las finanzas.

En este sentido, las finanzas han sido un lugar común donde diversos aprendizajes máquina pueden ser aplicados para solucionar un viejo problema, conocer el futuro. Es por ello, que en este trabajo se abordará una combinación de técnicas ya conocidas, como los son los indicadores bursátiles, algunos métodos de aprendizaje máquina, sus respectivos métodos de evaluación de desempeño y un conocimiento general de la situación económica mundial, para poder predecir los movimientos futuros de los índices accionarios en cuestión a un día.

El capítulo 1 hace una breve revisión de los tipos de aprendizaje máquina que existen y sus respectivas clases. De igual manera se hace una revisión profunda de la teoría sobre diferentes modelos de aprendizaje máquina para clasificación que se han de utilizar a lo largo de este trabajo, como son los métodos de clasificación bayesiana, modelos lineales generalizados, modelos lineales regularizados, Máquinas de Soporte Vectorial, Bosques Aleatorios y Vecinos más cercanos.

En el capítulo 2 se hace un recorrido sobre las formas en las que se puede evaluar el desempeño de los modelos de clasificación, empezando a través de los esquemas de validación cruzada, la calidad de los clasificadores y cómo medirlos de acuerdo a su clase y finalmente la comparación entre los diferentes clasificadores a través de pruebas paramétricas y pruebas no paramétricas. De tal manera de que con toda esta información se puedan seleccionar los diferentes hiperpámetros para los diversos modelos a través de validación cruzada, como eventualmente seleccionar el modelo con el mejor comportamiento frente a los datos.

El capítulo 3 habla de manera general sobre los índices financieros, cuáles son sus diferentes clasificaciones y características, pasando por las características particulares de los tres índices con los que estaremos trabajando (IPC, INMEX y S&P 500), su composición, un poco de su historia y la forma de su cálculo. Adicionalmente, se hace una breve reseña sobre los eventos más importantes que afectaron al mercado accionario en los últimos años, con la finalidad de tener un contexto alrededor del comportamiento de los movimientos de los índices accionarios.

El capítulo 4 hace un recorrido sobre los métodos de análisis de tendencia y predicción de movimientos bursátiles tradicionales, es decir, el análisis técnico. Esto se hace no solamente explicando el método gráfico de velas, sino también la clasificación, construcción y uso de los diversos indicadores técnicos que pueden obtenerse a través de la información bursátil, proporcionada en este caso por los índices. Algunos de estos indicadores posteriormente se han de utilizar para explicar y pronosticar el movimiento de los índices bursátiles elegidos para este trabajo.

En este trabajo se pretende definir si hubo un crecimiento o decrecimiento en el comportamiento del precio de los índices financieros; tomando en cuenta las métricas de comportamiento que se han visto a lo largo del capítulo 2, aplicadas a las técnicas de aprendizaje máquina vistas en el capítulo 1. Cabe mencionar que en este trabajo no se tomará en cuenta el tiempo explícitamente para el análisis del comportamiento de los datos. Lo anterior debido a que el tiempo se incluye de manera implícita en la construcción de los indicadores técnicos y la estructura que éstos manejan per-se. De tal manera que no es necesario analizar estos datos con series de tiempo.

En el capítulo 5 se hace una descripción general de los históricos de los índices accionarios, la arquitectura de la información, es decir, los parámetros e hiperparámetros que se han de ocupar para las formas de evaluación vistas en el capítulo 2; y finalmente la ejecución de las técnicas de aprendizaje máquina vistas en el capítulo 1. Además de evaluar los resultados finales para poder seleccionar el modelo que haga una mejor predicción en los movimientos de los diferentes índices bursátiles.

Capítulo 1

Técnicas de clasificación en aprendizaje de máquina

1.1. Introducción

Este capítulo tiene la intención de estudiar algunas de las técnicas más populares utilizadas en el aprendizaje de máquina aplicado a clasificación. Aunque los tipos de aprendizaje que se tienen dentro de la ciencia de datos, son el aprendizaje supervisado, el no supervisado y el semi-supervisado, este capítulo únicamente abarcará al aprendizaje supervisado.

Además se ahondará a mayor profundidad el estudio de las siguientes técnicas o modelos de clasificación: regresión Logit y Probit, regresión regularizada Ridge y Lasso, máquinas de soporte vectorial (MSV), clasificación vía Naive-Bayes, Vecinos más cercanos (KNN) y Bosques Aleatorios.

1.2. Tipos de aprendizaje en ciencia de datos

La mayoría de los problemas de aprendizaje estadístico se pueden clasificar dentro de algunas de las siguientes tres categorías, aprendizaje supervisado, aprendizaje no supervisado y aprendizaje semisupervisado.

1.2.1. Aprendizaje supervisado

En el caso supervisado, para cada observación de la predicción a medir $x_i,\ i = 1, \ldots, n$ existe una variable de respuesta asociada y_i. De esta manera se ajusta un modelo que relacione la respuesta con las variables explicativas, con el objetivo de poder predecir con precisión las respuestas para observaciones futuras. Los métodos de aprendizaje estadístico, tales como la Regresión Lineal, la Logística, clasificador Naive Bayes, las máquinas de

soporte vectorial, Bosques Aleatorios, etc; tienen como base el aprendizaje supervisado. Adicionalmente, el aprendizaje supervisado puede ser separado en dos grandes grupos, la clasificación y la regresión.

- Clasificación. Esta categoría utiliza un algoritmo para asignar con precisión datos de prueba en categorías específicas. Reconoce características específicas dentro del conjunto de datos utilizado e intenta obtener algunas conclusiones sobre cómo estas características deberían de ser etiquetadas. Algunos ejemplos de algoritmos de clasificación son: máquinas de soporte vectorial, Árboles de Decisión, Bosques Aleatorios, etc.

- Regresión. Estos algoritmos se utilizan para comprender la relación entre las variables explicativas y las variables de respuesta, suelen ser usadas para proyecciones. La regresión lineal, la regresión logística y la regresión polinomial son algoritmos de regresión populares.

1.2.2. Aprendizaje no supervisado

Por otro lado, el aprendizaje no supervisado describe una situación más compleja. Para cada una de las observaciones $i = 1, \ldots, n$ se tiene un vector de medidas x_i pero ninguna respuesta asociada y_i. De tal manera que no es posible ajustar un modelo de regresión lineal, ya que no hay una variable de respuesta a predecir. De aquí se obtiene su nombre de "aprendizaje no supervisado" porque se carece de una variable de respuesta que pueda supervisar el análisis hecho. Algunos ejemplos de este tipo de aprendizaje son los múltiples métodos de clustering, como lo son: K vecinos cercanos, Análisis de Componentes Principales, Análisis de Componentes Independientes, etc.

1.2.3. Aprendizaje semi-supervisado

Naturalmente, la cuestión de si un análisis debería de considerarse como supervisado o bien como no supervisado llega a ser bastante confuso. Por ejemplo, supóngase que se tiene un conjunto de n observaciones, donde $m < n$, se tiene también variables de respuesta y variables explicativas. Para las $m - n$ observaciones, se tiene una medición, pero no una respuesta. Este escenario puede surgir siempre y cuando las variables de explicativas puedan ser medidas de manera barata, no así las variables de respuesta. A este escenario se le conoce como un problema de aprendizaje semi-supervisado.

Este método de aprendizaje puede incorporar las m observaciones para las que disponen de variables de respuesta, así como las $n - m$ observaciones que no las tienen.

1.3. Un poco de teoría de clasificadores

En aprendizaje estadístico, el problema de clasificación se considera un método de aprendizaje supervisado, i.e. inferir una función $f(x)$ a partir de datos etiquetados.

El conjunto de datos de entrenamiento consiste de un vector input, vector de features o vector de covariables $x = (x_1, \ldots, x_d)$; y un valor output, i.e., etiqueta de clase, ó simplemente etiqueta $y \in \{C_1, \ldots, C_K\}$.

Dado un conjunto de entrenamiento, la tarea del algoritmo de clasificación es analizar estos datos y generar una función que se use para clasificar nuevas observaciones (aún no observadas) y asignales una etiqueta a cada una de ellas.

Una subclase común de clasificación es la clasificación probabilista. Estos algoritmos encuentran la mejor clase para un vector de características dado, pero en lugar de simplemente asignar la mejor clase como otros algoritmos de clasificación, los algoritmos de clasificación probabilista también estimarán la probabilidad de que una observación pertenezca a cada una de las posibles clases. La clase con la probabilidad más alta, generalmente es la que se selecciona como la mejor clase.

En general, los algoritmos de clasificación tienen algunas ventajas con respecto a los clasficadores no-probabilistas, como son:

(i) Se puede obtener un intervalo de confianza asociado con la etiqueta de clase seleccionada, y por lo tanto abstenerse en caso de que la confianza sea muy baja.

(ii) Los clasificadores probabilísticos se pueden incorporar de manera más efectiva en tareas de aprendizaje estadísticos de gran escala, y por lo tanto se puede reducir el problema de error de propagación.

En un marco probabilista, el punto principal de la clasificación probabilista es estimar la probabilidad de clase posterior $\mathbb{P}(C_k|x)$. Después de obtener dichas probabilidades posteriores, se usa teoría de decisión para determinar la clase a la que pertenece cada nuevo input x.

De manera muy general, hay dos formas en las que se puede estimar la probablidad posterior de clase.

En la primera forma, el objetivo es determinar las probabilidades condicionales de clase $\mathbb{P}(x|C_k)$ para cada clase individualmente, e inferir las prior de clase $\mathbb{P}(C_k)$. De manera equivalente, se puede modelar la distribución conjunta $\mathbb{P}(x, C_k)$ directamente y posteriormente normalizarla para obtener las probabilidades posteriores. Como las probabilidades condicionales de clase definen el proceso aleatorio que genera las covariables que se miden,

estas metodologías que modelan implícita o explícitamente la distribución de los inputs y los outputs se conocen como modelos generativos. Si los datos observados realmente se muestrean del modelo generativo, entonces el ajuste de parámetros del modelo generativo generalmente se hace por máxima verosimilitud. Dos ejemplos clásicos de modelos generativos para clasificación son el clasificador Naive Bayes y el modelo de Markov oculto (Hidden Markov Model).

Otra clase de modelos es la que directamente modela las probabilidades posteriores $\mathbb{P}(C_k|x)$ mediante el aprendizaje de una función que discrimina $g(x) = \mathbb{P}(C_k|x)$, que relaciona directamente x con la estiqueta de clase C_k. Este tipo de metodología se conoce como modelo discriminativo. Por ejemplo, en el caso de un problema de clasificación binaria, $g(x) = \mathbb{P}(C_k|x)$ puede ser un valor continuo entre 0 y 1; tal que $f < \frac{1}{2}$ representa a la clase C_1 y $f \geq \frac{1}{2}$ representa a la clase C_2. Ejemplos de este tipo de modelos discriminiativos son la regresión logística y la regresión Probit.

1.4. Clasificación Bayesiana

Sean X, Y variables aleatorias, tales que Y es discreta. Recuérdese que a partir del Teorema de Probabilidad Total

$$\mathbb{P}(X = x) = \sum_{y \in sop(Y)} \mathbb{P}(X = x, Y = y) \tag{1.1}$$

y a partir de la definición de probabilidad condicional

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(Y = y|X = x)p(X = x) \tag{1.2}$$

que se conoce como regla del producto.

A partir de esta regla del producto y la propiedad de simetría $\mathbb{P}(X, Y) = \mathbb{P}(Y, X)$ se puede obtener el Teorema de Bayes

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)} \tag{1.3}$$

y en el caso particular de que Y sea discreta

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y)}{\sum_{y \in sop(Y)} \mathbb{P}(X = x|Y = y)p(Y = y)} \tag{1.4}$$

El denominador de la fórmula de Bayes se puede considerar como una constante de normalización que se requiere para que $\mathbb{P}(Y|X)$ sea efectivamente una función de probabilidad.

Ejemplo 1.4.1.

Supóngase que se tienen dos cajas, una roja y otra blanca. Dentro de la caja roja hay 2 manzanas, 4 limones y 6 naranjas. Dentro de la caja blanca hay 3 manzanas, 6 limones y 1 naranja. Supóngase que aleatoriamente se selecciona una de las dos cajas y de dicha caja se selecciona aleatoriamente una fruta. Se observa, se registra el resultado y se devuelve a la caja de donde vino. Este proceso se repite varias veces. Supóngase que el 40 % de las veces se seleccionó la caja roja y el 60 % la caja blanca. Sea Y la variable aleatoria para la selección de la caja. Sea r la caja roja y b la caja blanca, entonces si m son las manzanas, l los limones y n las naranjas, se tiene que

$$\mathbb{P}(Y = r) = \frac{4}{10}, \quad \mathbb{P}(Y = b) = \frac{6}{10}$$

Sea X la variable aleatoria para la selección de la fruta dentro de la caja. Entonces

$$\mathbb{P}(X = m|Y = r) = \frac{2}{12},\ \mathbb{P}(X = l|Y = r) = \frac{4}{12},\ \mathbb{P}(X = n|Y = r) = \frac{6}{12}$$

$$\mathbb{P}(X = m|Y = b) = \frac{3}{10},\ \mathbb{P}(X = l|Y = b) = \frac{6}{10},\ \mathbb{P}(X = n|Y = b) = \frac{1}{10}$$

Supóngase que se seleccionó una fruta, ésta resultó ser una naranja y se quiere saber de qué caja proviene. A partir de la fórmula de Bayes,

$$\mathbb{P}(Y = r|X = n) = \frac{\mathbb{P}(X = n|Y = r)\mathbb{P}(Y = r)}{\mathbb{P}(X = n)}$$

$$= \frac{\mathbb{P}(X = n|Y = r)\mathbb{P}(Y = r)}{\mathbb{P}(X = n|Y = r)\mathbb{P}(Y = r) + \mathbb{P}(X = n|Y = b)\mathbb{P}(Y = b)}$$

$$= \frac{\frac{6}{12}\frac{4}{10}}{\frac{6}{12}\frac{4}{10} + \frac{1}{10}\frac{6}{10}} = \frac{10}{13}$$

y por complementos $\mathbb{P}(Y = b|X = n) = 1 - \frac{10}{13} = \frac{3}{13}$.

▽

En general, se tiene interés en probabilidades de clases dadas observaciones muestrales. Supóngase que se usa a la variable aleatoria Y para denotar a la etiqueta de clase de las observaciones y al vector aleaorio X para representar a las (co)variables de las observaciones.

Se puede interpretar a $\mathbb{P}(Y = C_k)$ como la probabilidad prior para la clase k, que representa la probabilidad de que la etiqueta de clase de un punto sea C_k antes de observar sus características. Una vez que se observaron las características X de los datos muestrales, se puede utilizar el Teorema de Bayes para calcular a la probabilidad posterior $\mathbb{P}(Y|X)$. La cantidad $\mathbb{P}(X|Y)$ se puede interpretar como qué tan probable es el dato observado para

las diferentes clases, que se conoce como verosimilitud.

En este sentido se puede interpretar a la fórmula de Bayes

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)}{p(X)} p(Y) \tag{1.5}$$

cómo un mecanismo que va de $\mathbb{P}(Y)$ a $\mathbb{P}(Y|X)$ aprendiendo de los datos mediante el factor $\frac{\mathbb{P}(X|Y)}{\mathbb{P}(X)}$

1.4.1. Clasificador Naive-Bayes

El clasificador Naive-Bayes se conoce como el clasificador bayesiano más simple, que se ha convertido en un importante modelo probabilístico, muy exitosos en aplicaciones a pesar de su fuerte suposición de independencia condicional.

Supóngase que se tiene un conjunto de datos de entrenamiento $\{(x_i, y_i)\}_{i=1}^n$, donde cada x_i es un vector d-dimensional que contiene a las características, covariables o predictores de la observación i y y_i denoto la etiqueta de clase de la observación i. Se supondrán variables aleatorias Y y X con componentes $X_1, \ldots, X_d$ correspondientes a la etiqueta y y al vector de características $x = (x_1, \ldots, x_d)$. En general, Y es una variable aleatoria discreta que toma los valores de las posibles etiquetas de clases $C_1, \ldots, C_K$ y las características $X_1, \ldots, X_d$ pueden ser discretas ó contiuas.

Una de las tareas principales es entrenar un clasificador que devuelva la probabilidad posterior $\mathbb{P}(Y|X)$ para todos los posibles valores de Y.

En virtud del Teorema de Bayes, se puede escribir a $\mathbb{P}(Y = C_k|X = x)$ de la siguiente manera.

$$\begin{aligned} \mathbb{P}(Y = C_k|X = x) &= \frac{\mathbb{P}(X = x|Y = C_k)\mathbb{P}(Y = C_k)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(X_1 = x_1, \ldots, X_d = x_d|Y = C_k)\mathbb{P}(Y = C_k)}{\mathbb{P}(X_1 = x_1, \ldots, X_d = x_d)} \end{aligned} \tag{1.6}$$

Una manera de aprender $\mathbb{P}(Y|X)$ es usar a los datos de entrenamiento para estimar $\mathbb{P}(X|Y)$ y $\mathbb{P}(Y)$; y después usar dichos estimados junto con el Teorema de Bayes para determinar $\mathbb{P}(Y|X = x_j)$ para cualquier nuevo vector de predictores x_j.

En general, no es sencillo encontrar clasificadores bayesianos exactos, pues suelen ser computacionalmente demandantes e incluso intratables. Para ver esto, considerese el siguiente ejemplo.

Ejemplo 1.4.2.

Supóngase que Y es una variable booleana y X es un vector de dimensión d de predictores booleanos. Para cualquier C_k hay 2^d posibles valores de x, por lo tanto se necesita estimar $2(2^d - 1)$ parámetros $\mathbb{P}(X_1 = x_1, \ldots, X_d = x_d | Y = C_k)$. Por ejemplo, si X fuera un vector con 20 predictores booleanos, se tendría que estimar 1,048,576 de parámetros.

Para manejar esta complejidad poco tratable, el clasificador Naive-Bayes reduce esta complejidad al hacer una suposición de independencia condicional de los predictores $X_1, \ldots, X_d$, i.e. supone que dado Y, las variables $X_1, \ldots, X_d$ son independientes.

Para el ejemplo 4.2, la suposición de independencia condicional reduce dramáticamente el número de parámetros a estimar para modelar $\mathbb{P}(X|Y)$ de $2(2^d - 1)$ a simplemente $2d$.

Esta suposición de independencia condicional, lleva a que

$$\mathbb{P}(X_1 = x_1, \ldots, X_d = x_d | Y = C_k) =$$

$$\prod_{j=1}^{d} \mathbb{P}(X_j = x_j | X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, Y = C_k)$$

$$\prod_{j=1}^{d} \mathbb{P}(X_j = x_j | Y = C_k) \tag{1.7}$$

Esta es la suposición Naive-Bayes y es relativamente fuerte aunque muy útil.

Con la suposición Naive-Bayes, se puede reescribir a la expresión (1.6) como

$$\begin{aligned} \mathbb{P}(Y = C_k | X_1, \ldots, X_d) &= \frac{\mathbb{P}(Y = C_k) \prod_j \mathbb{P}(X_j | Y = C_k)}{\sum_i \mathbb{P}(Y = C_i) \prod_j \mathbb{P}(X_j | Y = C_i)} \\ &= \frac{\mathbb{P}(Y = y_k) \prod_j \mathbb{P}(X_j | Y = y_k)}{\sum_i \mathbb{P}(Y = y_i) \prod_j \mathbb{P}(X_j | Y = y_i)} \end{aligned} \tag{1.8}$$

Si sólo se tiene interés en el valor más probable de Y, entonces se tiene la siguiente regla de clasificiación Naive-Bayes

$$Y \leftarrow arg\,\text{máx} \left\{ C_k \frac{\mathbb{P}(Y = C_k) \prod_j \mathbb{P}(X_j | Y = C_k)}{\sum_i \mathbb{P}(Y = C_i) \prod_j \mathbb{P}(X_j | Y = C_i)} \right\} \tag{1.9}$$

Como el denominador de la función objetivo no depende de C_k, esta formulación se puede simplificar como

$$Y \leftarrow arg\,\text{máx} \left\{ C_k \mathbb{P}(Y = C_k) \prod_j \mathbb{P}(X_j | Y = C_k) \right\} \tag{1.10}$$

1.4.2. Estimación máximo verosímil para modelos Naive-Bayes

El modelo Naive-Bayes tiene dos tipos de parámetros que se deben estimar. El primero de ellos es

$$\pi_k = \mathbb{P}(Y = C_k) \tag{1.11}$$

para todos los valores C_k de Y. Este parámetro se puede interpretar como la probabilidad de observar a la etiqueta C_k. En este caso se tienen las restricciones de que $\pi_1, \ldots, \pi_K \geq 0$ y $\pi_1 + \ldots + \pi_K = 1$. Por lo tanto sólo se tienen $K - 1$ parámetros libres.

Para los predictores X_i supóngase que cada uno de ellos puede tomar J posibles valores discretos, entonces el segundo tipo de parámetros a estimar son

$$\theta_{ijk} := \mathbb{P}(X_i = x_{ij} | Y = C_k) \tag{1.12}$$

para cada predictor X_i, cada uno de sus posibles valores x_{ij} y cada uno de los posibles valores de C_k.

Observación 1.4.1.

En el supuesto de que los predictores X_i sean continuos, entonces se deberán de discretizar en n intervalos, donde cada intervalo fungirá como una clase distinta. En el proceso de Cross-Validation, se estimará en cuántas n particiones se deberá de dividir a los predictores X_i continuos para un mejor desempeño.

∇

El valor de θ_{ijk} se puede interpretar como la probabilidad de que la variable X_i tome el valor x_{ij} condicionado a que la etiqueta subyacente es C_k.

Obsérvese que para cualesquiera i, k, se debe satisfacer que $\sum_j \theta_{ijk} = 1$ y por lo tanto habrá $d(J-1)K$ parámetros libres.

Estos parámetros se pueden estimar usando estimadores máximo verosímiles basados en las frecuencias relativas en los datos. Los estimadores máximo verosímiles para θ_{ijk} dado un conjunto de datos de entrenamiento son

$$\hat{\theta}_{ijk} = \hat{\mathbb{P}}(X_i = x_{ij}|Y = C_k) = \frac{\text{conteo}(X_i = x_{ij}, Y = C_k)}{\text{conteo}(Y = C_k)} \tag{1.13}$$

Para evitar el caso en el que no hay observaciones en el numerador, es común adaptar un estimado suavizado

$$\hat{\theta}_{ijk} = \hat{\mathbb{P}}(X_i = x_{ij}|Y = C_k) = \frac{\text{conteo}(X_i = x_{ij}, Y = C_k) + l}{\text{conteo}(Y = C_k) + l \cdot J}, \tag{1.14}$$

donde J es el número de valores distintos que puede tomar X_i y l es un parámetro que modula el grado de suavizamiento. Si $l = 1$, este método se conoce como suavizamiento de Laplace.

Los estimados máximo verisímiles para π_k se pueden tomar como

$$\hat{\pi}_k = \hat{p}(Y = C_k) = \frac{\text{conteo}(Y = C_k)}{n}, \tag{1.15}$$

donde n es el número de observaciones en el conjunto de entrenamiento.

1.5. Modelos lineales generalizados

1.5.1. Regresión Logit

El modelo lineal generalizado es la unificación de los modelos lineales y no lineales de regresión, lo que permite la incorporación de distribuciones no normales. Dentro de los modelos lineales generalizados, se debe de cumplir que las distribuciones de las variables de respuesta sean de la forma exponencial, en otras palabras, pertenezca a la familia exponencial. Dentro de esta familia de distribuciones tenemos a la normal, binomial, exponencial, etc.

Si se usa esta aproximación para predecir una respuesta positiva, entonces se obtiene un modelo cuya variable de respuesta Y toma valores negativos para valores en X muy cercanos al cero, así como valores mayores a 1 para valores en X muy grandes.

En principio, siempre se pueden predecir $\mathbb{P}(X) < 0$ para algunos valores de X y $\mathbb{P}(X) > 1$ para otros, si y solo si, el rango de valores de X no esta acotado. Para evitar este tipo de problemas, se debe modelar a $\mathbb{P}(X)$ usando una función cuyos valores de salida se encuentren entre el 0 y el 1 para todos los valores de X. Muchas funciones cumplen esta condición, sin embargo, para las regresiones logísticas se usará la función logística.

La regresión logística forma parte de los modelos lineales generalizados, y es usada para modelar la probabilidad de cierta clase de eventos donde la variable de respuesta es binaria,

es decir, cae en dos categorías tales como: si o no; sobrevivir o morir; ó éxito o fracaso. Normalmente se denotaría como un 0 o 1, donde el cero es el fracaso y el uno se define como éxito. Cabe señalar que la respuesta es esencialmente cualitativa, pues la designación de éxito o fracaso es arbitraria.

Supóngase que el modelo tiene la siguiente forma:

$$y_1 = x_1'\beta + \epsilon_i \tag{1.16}$$

en donde $x_1' = [1, x_{i1}, ..., x_{ik}]$, $\beta' = [\beta_0, \beta_1, ..., \beta_k]$, y la variable de respuesta y_j toma el valor entre 0 y 1, por lo tanto, se supondría que la variable y_i es una variable aleatoria que se distribuye Bernoulli con probabilidad 1 si $\mathbb{P}(y_i = 1) = \pi_i$ y 0 si $\mathbb{P}(y_i = 0) = 1 - \pi_i$.

Además, si la $\mathbb{E}(\epsilon_i) = 0$ la esperanza de la variable de respuesta es

$$\mathbb{E}(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{1.17}$$

por lo tanto,

$$\mathbb{E}(y_i) = x_i'\beta = \pi_i. \tag{1.18}$$

Lo que significa que la variable de respuesta dada por la función $\mathbb{E}(y_t) = x_t'\beta$ es la probabilidad de que la variable de respuesta tome el valor de 1.

Hay que tener en cuenta algunos detalles con respecto a este modelo de regresión, el primero es que debido a que la respuesta es binaria, entonces los términos del error ϵ_t solo puede tomar dos valores, i.e.,

$$\epsilon_t = 1 - x_t'\beta \quad \text{donde} \quad y_t = 1 \epsilon_t = -x_t'\beta \quad \text{donde} \quad y_t = 0 \tag{1.19}$$

Debido a esto, los errores de este modelo no pueden ser normales.

De igual manera, el error de la varianza no es constante, ya que,

$$\sigma^2_{y_i} = \mathbb{E}\left[y_t - \mathbb{E}(y_t)\right]^2 = (1 - \pi_t)^2\pi_t + (0 - \pi_t)^2(1 - \pi_t) = \pi_t(1 - \pi_t) \tag{1.20}$$

Vease que la expresión anterior queda resumida como:

$$\sigma^2_{y_t} = \mathbb{E}(y_t)[1 - \mathbb{E}(y_t)] \tag{1.21}$$

tomando en cuenta que $\mathbb{E}(y_t) = x_t'\beta = \pi_t$. Esto indica que la varianza de las observaciones, que es la misma que la varianza del error, es una función de la media.

Finalmente, hay una restricción en la función de respuesta, debido a que $\mathbb{E}(y_t)$ es una probabilidad,

$$0 \leq \mathbb{E}(y_t) = \pi_t \leq 1 \tag{1.22}$$

Esta restricción puede causar problemas con la elección de la función lineal de respuesta. Es posible ajustar un modelo con los datos donde los valores predichos se soporten fuera de los valores 0 y 1.

Cuando se trata con variables de respuesta binarias, existe suficiente evidencia que indica que la forma de la función de respuesta sería no lineal. Una curva monótona creciente en forma de S o S invertida podría ser utilizada para dicho fin. Esta función es conocida como la función de respuesta logística y tiene la forma

$$\mathbb{E}(y) = \frac{exp(x'\beta)}{1 + exp(x'\beta)} = \frac{1}{1 + exp(-x'\beta)} \tag{1.23}$$

Ahora con esta función, nótese que para valores en X pequeños o cercanos al cero, la probabilidad se vuelve cercana al cero, pero nunca se encuentra por debajo del cero. De la misma manera, para valores muy grandes en X la probabilidad de éxito o fracaso se vuelve muy cercana a uno, pero nunca sobrepasa el uno. La función logística siempre produciría una curva en forma de "S", y sin importar los valores de X, se obtendría una predicción sensible.

Sin embargo, la función de respuesta puede ser transformada en una función lineal. Una aproximación puede ser a través de definir una porción del modelo en términos de la función de respuesta de la media, es decir, sea

$$\eta = x'\beta \tag{1.24}$$

la variable de predicción donde η se define por la siguiente transformación

$$\eta = \ln\left(\frac{\pi}{1-\pi}\right) \tag{1.25}$$

A esta transformación se le conoce como transformación **Logit** de probabilidad π, y el ratio $\pi/(1-\pi)$ en la transformación se le conoce como odds o bien como momios.

Cabe mencionar que a diferencia de una regresión lineal, en una regresión logística el incremento de X por una unidad de cambio en los log-odds por β_1 o su equivalente se multiplica por el odds de e^{β_1}. Sin embargo, debido a esta relación entre $\mathbb{P}(X)$ y X no es una recta pues β_1 no corresponde al cambio de $\mathbb{P}(X)$ asociado a una unidad de incremento en X. La cantidad que $\mathbb{P}(X)$ cambia debido a una unidad de cambio en X dependería del valor actual de X.

1.5.1.1. Estimación de parámetros de la regresión logística

La forma general de la regresión logística es

$$y_t = \mathbb{E}(y_t) + \epsilon_t \tag{1.26}$$

donde las observaciones y_t son variables aleatorias independientes Bernoulli cuya esperanza es

$$\mathbb{E}(y_t) = \pi_t = \frac{exp(x_t'\beta)}{1 + exp(x_t'\beta)} \tag{1.27}$$

Después se usa el método de máxima verosimilitud para estimar los parámetros de la regresión lineal $x_t'\beta$.

Cada observación muestra sigue la distribución Bernoulli, por lo que la probabilidad de cada distribución para cada muestra se define como

$$f_t(y_t) = \pi_t^{y_t}(1 - \pi_t)^{1-y_t}, \quad i = 1, ..., n \tag{1.28}$$

y cada observación y_t toma el valor de 0 y 1. Debido a que las observaciones son independientes, la verosimilitud funciona como producto de las funciones individuales, es decir,

$$L(y_1, ..., y_n, \beta) = \prod_{t=1}^{n} f_t(y_t) = \prod_{t=1}^{n} \pi_i^{y_t}(1 - \pi_t)^{1-y_t} \tag{1.29}$$

aplicando la función logaritmo a la igualdad anterior, se convertiría en una log-verosimilitud, que es representada como

$$\ln\left(L(y_1, ..., y_n, \beta)\right) = \ln \prod_{t=1}^{n} f_t(y_t) = \sum_{t=!}^{n} \left[y_t \ln\left(\frac{\pi}{1 - \pi_t}\right)\right] + \sum_{t=1}^{n} \ln(1 - \pi_t) \tag{1.30}$$

como $(1 - \pi_t) = [1 + exp(x_t'\beta)]^{-1}$ y $\eta_t = \ln[\pi_t/(1 - \pi_t)] = x_t'\beta$, entonce se llega a que la log-verosimilitud se puede escribir como

$$\ln(L(y, \beta)) = \sum_{t=1}^{n} y_t x_t'\beta - \sum_{t=1}^{n} \ln[1 + exp(x_t'\beta)] \tag{1.31}$$

Si se toma a y_t como el resultado de que la variable tome el valor de 1 para la i-iésima observación y η_t como el número de intentos para cada observación. Entonces la log-verosimilitud se convierte en

$$\begin{aligned} \ln L(y,\beta) = \sum_{t=1}^{n} y_t \ \ln(\pi_t) \ + \ \sum_{t=1}^{n} n_t \ln(1-\pi_t) - \sum_{t=1}^{n} y_t \ln(1-\pi_t) = \\ \sum_{t=1}^{n} y_t \ \ln(\pi_t) + \sum_{t=1}^{n} (n_t - y_t) \ln(1-\pi_t) \end{aligned} \tag{1.32}$$

Sea $\hat{\beta}$ el parámetro estimado, si las suposiciones del modelo son correctas, entonces se puede mostrar asintoticamente que

$$\mathbb{E}(\hat{\beta}) = \beta \ \text{ y } \ \text{Var}(\hat{\beta}) = (X'VX)^{-1} \tag{1.33}$$

donde la matriz diagonal V de n x n contiene la varianza estimada de cada observación en la diagonal principal, que es el i-iésimo elemento de la diagonal V, es decir,

$$V_{tt} = n_t \hat{\pi}_t (1 - \hat{\pi}_t) \tag{1.34}$$

El valor del estimador lineal es $\eta_t = x_t' \hat{\beta}$ y el valor ajustado de la regresión logística se escribe como

$$\hat{y_t} = \hat{\pi}_t = \frac{exp(\hat{\eta}_t)}{1 + exp(\hat{\eta}_t)} = \frac{exp(x_t'\hat{\beta})}{1 + exp(x_t'\hat{\beta})} = \frac{1}{1 + exp(-x_t'\hat{\beta})} \tag{1.35}$$

1.5.1.2. Interpretación de los parámetros en una regresión lineal logística

Es fácil interpretar los parámetros de una regresión logística, considérese el caso donde el estimador lineal tiene un solo estimador, por lo tanto, el valor ajustado al estimador lineal en un valor particular de x, digamos x_t es

$$\hat{\eta}(x_t) = \hat{\beta}_0 + \hat{\beta}_t x_t \tag{1.36}$$

el valor ajustado en $x_t + 1$ es

$$\hat{\eta}(x_t + 1) \ = \ \hat{\beta}_0 \ + \ \hat{\beta}_1 (x_t + 1) \tag{1.37}$$

y la diferencia en dos valores estimados es

$$\hat{\eta}(x_t + 1) - \hat{\eta}(x_t) \ = \ \hat{\beta}_t \tag{1.38}$$

Ahora, $\hat{\eta}(x_t)$ se conocerá como log-odds cuando la variable de regresión sea igual a x_t y $\hat{\eta}(x_t + 1)$ es igual al log-odds cuando la variable de regresión sea $x_t + 1$. Por lo tanto la diferencia entre los dos ajustes es

$$\hat{\eta}(x_t+1) - \hat{\eta}(x_t) = \ln(\text{odds}_{\text{x}_\text{t}+1} - \text{odds}_{\text{x}_\text{t}}) = \ln\left(\frac{\text{odds}_{\text{x}_\text{t}+1}}{\text{odds}_{\text{x}_\text{t}}}\right) = \hat{\beta}_1 \tag{1.39}$$

Si aplicamos la exponencial a ambos lados de la última igualdad obtenemos

$$\hat{O}_R = \frac{\text{odds}_{\text{x}_\text{t}+1}}{\text{odds}_{\text{x}_\text{t}}} = e^{\hat{\beta}_1} \tag{1.40}$$

Los ratio odds pueden ser interpretados como los estimadores crecientes de la probabilidad de éxito asociados con una unidad de cambio en el valor del estimador. En general, la estimación que incrementa en el ratio asociado con el cambio de d unidades en el estimador es $exp(d\hat{\beta}_1)$.

Cabe resaltar que existe una similitud entre los ratio odds en regresión logística y la tabla de contingencia de 2x2 que es usada en el análisis de información categórica.

La inferencia estadística en los modelos de regresión logística están basados en las propiedades de máxima verosimilitud de los estimadores y en la verosimilitud de las pruebas de ratio. Es decir, muestras grandes o resultados asintóticos.

1.5.1.3. Prueba de razón de verosimilitud

Una prueba de razón de verosimilitud (RV) puede ser usada para comparar un modelo completo a través de un modelo reducido que sea de interés. Estas pruebas comparan dos veces el algoritmo del valor de la función de verosimilitud para el modelo completo (MC) y dos veces el algoritmo del valor de la función de verosimilitud del modelo reducido (MR) para obtener la prueba estadística,

$$RV = 2ln\frac{L(MC)}{L(MR)} = 2[\ln L(FM) - \ln L(RM)] \tag{1.41}$$

Para grandes muestras, cuando se hace la reducción del modelo de manera correcta, la prueba estadística RV sigue una distribución Ji-Cuadrada con grados de libertad igual a la diferencia en el número de parámetros entre el modelo completo y el reducido. Por lo tanto, si la prueba estadística RV excede el porcentaje α de puntos de esa distribución Ji-Cuadrada, por ende se rechaza el hecho de que dicho modelo reducido sea apropiado.

La razón de verosimilitud aproximada puede ser usada para pruebas de significancia en regresiones logísticas. Esta prueba usa el modelo que ha sido ajustado a los datos del

modelo completo y lo compara con un modelo reducido que tiene alta probabilidad de éxito. Este modelo de probabilidad constante de éxito es

$$\mathbb{E}(y) = \pi = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{1.42}$$

que es, el modelo de regresión logística sin variables regresoras. La máxima verosimilitud estimada de la probabilidad constante de éxito es sólo y/n, donde y es el número total de éxitos y n es el número de observaciones. Substituyendo esto en la función de log-verosimilitud, se da como resultado el máximo valor para la función log-verosimilitud del modelo reducido como

$$\ln\ L(MR) = y\ \ln(y) + (n-y)\ln(n-y) - n\ln(n) \tag{1.43}$$

Por lo tanto, la prueba estadística de razón de verosimilitud para pruebas de significancia de una regresión es

$$LR = 2\left[\sum_{t=1}^{n} y_t \ln(\hat{\pi_t}) + \sum_{t=1}^{n}(n_t - y_t)\ln(1-\hat{\pi_t}) - [y\ln(y) + (n-y)\ln(n-y) - n\ln(n)]\right] \tag{1.44}$$

Un valor grande, como resultado de esta prueba podría indicar que al menos uno de las variables explicativas en la regresión logística es importante porque tienen un coeficiente distinto del cero.

1.5.1.4. Prueba de bondad de ajuste

Las pruebas de bondad de ajuste para un modelo de regresión logística pueden ser calificadas usando una prueba de razón de verosimilitud. Esta prueba compara el modelo actual con un modelo saturado (MS), donde cada observación o grupo de ellas, se le permite tener su propio parámetro. Este parámetro o probabilidad de éxito es y_t/n_t donde y_t es el número de éxitos y n_t es el número total de observaciones. La desviación se define como el doble de la diferencia entre la log-verosimilitud entre la saturación del modelo y el modelo completo, el cual suele ser el modelo en si, que ha sido ajustado a los datos con probabilidad ajustada de éxito $\hat{\pi_t} = exp(x_t'\hat{\beta})/\left[1 + exp(x_t'\hat{\beta})\right]$. La desviación se define como

$$D = 2\ln\frac{L(MS)}{L(FM)} = 2\sum_{t=1}^{n}\left[y_t \ln\left(\frac{y_t}{n_t\pi_t} + (n_t - y_t)\ln(\frac{n_t - y_t}{n_t(1-\hat{\pi_t})})\right)\right] \tag{1.45}$$

Para hacer el cálculo de la desviación, nótese que $y\ln(y/n\hat{\pi}) = 0$, y si $y = n$ se tiene $(n-y)\ln\left[(n-y)/n(1-\hat{\pi}\right] = 0$. Cuando el modelo de regresión logística está ajustado

adecuadamente a los datos y el tamaño de muestra es grande, la desviación tiene una distribución Ji-Cuadrada con $n-p$ grados de libertad, donde p es el número de parámetros del modelo. Por otro lado, valores pequeños de la desviación (o números grandes de p-value) implican que el modelo posee un ajuste satisfactorio en los datos, mientras que valores grandes de desviación, implican que el modelo actual no es adecuado. Una regla que se puede ocupar aquí es dividir la desviación por el número de grados de libertad. Si la razón $D/(n-p)$ es mucho mayor que la unidad, el modelo no es adecuado para ajustar los datos.

La desviación tiene un análogo en la teoría de la regresión lineal normal. En el modelo de regresión lineal $D = SS_{\text{Res}}/\sigma^2$. Esta cantidad tiene una distribución Ji-Cuadrada con $n-p$ grados de libertad, si las observaciones son normales e independientes. Empero, la desviación en la teoría normal para regresiones lineales contienen el parámetro desconocido σ^2, por lo tanto no se puede calcular de manera directa. Sin embargo, a pesar de este detalle, la desviación y el residual de la suma de cuadrados son esencialmente equivalentes.

Al realizar el ajuste, este se puede probar a través de la prueba estadística "Person Ji-Cuadrada", la cual compara las observaciones y las probabilidades esperadas de éxito y fracaso de cada grupo de observaciones. El número esperado de éxitos es $n_t\hat{\pi}_t$ y el número esperado de fallas es $n_t(1-\hat{\pi}_t) = 1, ..., n$. La prueba estadística de Person Ji-Cuadrada es

$$\chi^2 = \sum_{t=1}^{n} \left\{ \frac{(y_t - n_t\hat{\pi}_t)^2}{n_t\hat{\pi}_t} + \frac{[(n_t - y_t) - n_t(1-\hat{\pi}_t)]^2}{n_t(1-n_t\hat{\pi}_t)} \right\} = \sum_{t=1}^{n} \frac{(y_t - n_t\hat{\pi}_t)}{n_t\hat{\pi}_t(1-\hat{\pi}_t)} \tag{1.46}$$

La estadística resultante de la Ji-Cuadrada de Person puede ser comparada con una distribución Ji-Cuadrada con $n-p$ grados de libertad. Valores pequeños de esta estadística (o valores grandes del p-value) implican que el modelo se ajusta satisfactoriamente en los datos. La estadística de la Ji-Cuadrada de Person puede ser también dividida en el número de grados de libertad $n-p$ y la razón compararla con la unidad. Si la razón excede a la unidad, la bondad de ajuste del modelo es incuestionable.

Cuando no hay replicas en las variables explicativas, las observaciones pueden agruparse para aplicar una bondad de ajuste llamada prueba de Hosmer-Lemeshow. En este procedimiento las observaciones se clasifican en g grupos basados en la estimación de probabilidades de éxito. Generalmente, cerca de 10 son usados (cuando $g = 10$ los grupos son llamados deciles de riesgo) y las observaciones de éxito llamadas O_j y de fracaso $N_j - O_j$ son comparadas con las frecuencias esperadas de cada grupo, $N_j\bar{\pi}_j$ and $N_j(1-\bar{\pi}_j)$, donde N_j es el número de observaciones en el j-iésimo grupo y el promedio estimado de la probabilidad de éxito en el j-iésimo es $\bar{\pi}_j = \sum_{i \in \text{grupo}j} \hat{\pi}/N_j$. La estadística de Hosmer-Lemeshow es tan solo una prueba de ajuste de bondad Ji-Cuadrada de Person la cual compara frecuencias observadas y esperadas:

$$HL = \sum_{j=1)}^{n} \frac{(O_j - N_j\bar{\pi}_j)^2}{N_j\bar{\pi}_j(1-\bar{\pi}_j)} \tag{1.47}$$

Si el modelo de regresión logística ajustado es correcto, entonces la estadística HL sigue a una distribución Ji-Cuadrada con $g-2$ grados de libertad cuando el tamaño de la muestra es grande. Por lo tanto, valores grandes de HL implican que el modelo no es adecuado para ajustar los datos. También es útil el calcular la razón de la estadística Hosmer-Lemeshow para el número de grados de libertad $g-p$ cuyos valores, si son cercanos a la unidad implican un ajuste adecuado.

1.5.1.5. Pruebas de hipótesis en subconjuntos de parámetros usando desviación

También se puede usar la desviación para pruebas de hipótesis en subconjuntos de parámetros de los modelos; esto al usar la diferencia en las sumas de cuadrados de regresión (o error) para probar hipótesis similares en el caso del modelo de regresión lineal de error normal. Recordar que el modelo puede escribirse como

$$\eta = X\beta = X_1\beta_1 + X_2\beta_2 \tag{1.48}$$

donde el modelo completo tiene p parámetros, β_1 contiene $p-r$ de éstos parámetro, β_2 contiene r de éstos parámetros, y las columnas de las matrices X_1 y X_2 contienen las variables asociadas con estos parámetros.

La desviación del modelo completo puede ser denotada por $D(\beta)$. Supóngase que se quiere hacer la siguiente prueba de hipótesis

$$H_0 : \beta_2 = 0, \qquad H_1 : \beta_2 \neq 0 \tag{1.49}$$

Por lo tanto, el modelo reducido es

$$\eta = X_1\beta_1 \tag{1.50}$$

Habrá que ajustar el modelo reducido, y deja que $D(\beta_1)$ sea la desviación del modelo reducido. La desviación para el modelo reducido sería siempre más grande que la desviación del modelo completo, porque el modelo reducido contaría con menos parámetros. Sin embargo, si la desviación para el modelo reducido no es mucho mayor a aquella del modelo completo, entonces, esto indicaría que el modelo reducido es tan bueno ajustando, como lo es el modelo completo, por lo tanto es probable que los parámetros en β_2 sean iguales a cero. Y por ende, no se puede rechazar la hipótesis nula planteada. Sin embargo, si la diferencia de la desviación es grande, al menos uno de los parámetros de β_2 es probablemente diferente

de cero, y se puede rechazar la hipótesis nula. Formalmente la diferencia en la desviación es

$$D(\beta_2|\beta_1) = D(\beta_1) - D(\beta) \tag{1.51}$$

y esta cantidad tiene $n - (p - r) - (n - p) = r$ grados de libertad. Si la hipótesis nula es verdadera y n es larga, entonces la diferencia en la desviación tiene una distribución Ji-Cuadrada con r grados de libertad. Por ende, la prueba estadística y la decisión son

- Si $D(\beta_2|\beta_1) \geq \chi^2_{\alpha,r}$, entonces se rechaza la hipótesis nula
- Si $D(\beta_2|\beta_1) < \chi^2_{\alpha,r}$, entonces no se rechaza la hipótesis nula

Algunas veces la diferencia en la desviación $D(\beta_2|\beta_1)$ se le conoce como desviación parcial.

1.5.2. Regresión Probit

La única diferencia entre la regresión Logit y Probit es la relación que existe entre π_i como funcion no lineal de $x_i'\beta$. La distribución de y_i es Bernoulli, pues y_i es una variable binaria y no existe alguna otra posibilidad.

La regresión Probit especifica que π equivale a la función de distribución de una variable aleatoria normal estándar en $x_i'\beta$, en otras palabras es la probabiliad de que una variable aleatoria con distribución normal estándar sea menor que $x_i'\beta$. Si se asume que existen una utilidad aleatoria no observada que conduce a la decisión de cada elección individual; esa utilidad aleatoria puede ser especificada como una distribución normal y ser igual a $x_i'\beta$ más un término de error. En ese caso, el modelo Probit se sigue como resultado de asumir que un individuo hace una decisión de elegir el evento respuesta cuando una utilidad aleatoria excede el límite que depende de x_i. Entonces el modelo Probit surge naturalmente de una regresión lineal para la utilidad subyacente no observada.

1.5.2.1. Función de enlace

Es una función que enlaza a Y con la estimación de Y. En general, una función de enlance es $F(.)$, tal que $F(Y) = X\beta + \epsilon$. Las funciones de enlace ayudan cuando se tienen variables dicotómicas, es decir, que toman únicamente dos valores, y se quieren transformar a variables continuas.Supongase por simplicidad que la variable dicotomía toma los valores 0 y 1, por lo tanto se tomaría el intervalo de $[0, 1]$ como dominio para que a través de una función se llegue a la recta real. Una función que va de la recta real a valores entre 0 y 1 es la función de distribución normal Φ, la cual asigna a cualquier valor Z, $\Phi(Z) \in [0, 1]$.

Por lo tanto, se puede decir que

$$\begin{aligned} \text{Si} \quad Y &= \Phi(X\beta + \epsilon) \\ \Phi^{-1}(Y) &= X\beta + \epsilon \\ Y' &= X\beta + \epsilon \end{aligned} \tag{1.52}$$

Entonces la función de enlace es $F(Y) = \Phi^{-1}(Y)$. A esta función se le conoce como enlace Probit.

En un modelo Probit, el valor de $X\beta$ es tomado como valor Z de una distribución normal. Entre mayor sea el valor de $X\beta$ signifca que el evento es más probable que suceda.

Una diferencia notable entre una regresión lineal estándar, contra un modelo Probit es que debido a la forma que ésta tiene, el cambio que tiene Y cuando X cambia, no es lineal,es decir, se puede recordar que el modelo lineal tiene la forma de

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_n X_n \quad \Longrightarrow \frac{\delta Y}{\delta X_i} = \beta_i \tag{1.53}$$

Por otra parte, el comportamiento del modelo Probit es

$$\begin{aligned} Y &= \Phi(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n) \Longrightarrow \\ \frac{\delta Y}{\delta X_i} &= \beta_i \ \phi \ (\beta_0 + \beta_1 x_1 + ... + \beta_n x_n) \end{aligned} \tag{1.54}$$

En otras palabras, la razón de cambio en un modelo Probit no solamente depende de β_i, sino también de los valores de x_i y las demás variables en la ecuación. Por lo que si se quiere medir el impacto que tiene x_i en Y, se tienen que elegir los valores para todas las demás variables x_j.

1.5.3. Regresión regularizada

1.5.3.1. Regresión Ridge

Como se vió anteriormente, la suma de cuadrados residuales

$$\text{SCR} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \beta_j x_{ij})\right)^2 \tag{1.55}$$

La regresión Ridge es muy similar al método de mínimo cuadrados, sin embargo, los coeficientes son estimados de manera diferente. En particular la regresión Ridge estima los coeficientes $\hat{\beta^R}$ que son los que se han de minimizar,

$$\sum_{i=1}^{n}\left(Y_i-(\hat{\beta}_0+\sum_{j=1}^{p}\beta_j x_{ij})\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \;=\; \text{SCR}+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{1.56}$$

donde $\lambda \geq 0$ es un parámetro de ajuste que ha de ser determinado de manera independiente.

La regresión Ridge al igual que la estimación de mínimos cuadrados busca estimar los coeficientes que se ajustan a los datos al minimizar la SCR, sin embargo, el segundo término, i.e., $\lambda\sum_{j=1}^{p}\beta_j^2$, también llamada penalización de contracción, es pequeña cuando $\beta_1,...,\beta_p$ son cercanos a cero. Por lo que tienen el efecto de contracción las estimaciones de β_j hacia cero.

El parámetro de ajuste λ sirve como control de impacto entre estos dos términos en la estimación de coeficientes de regresión. De esta manera, si $\lambda = 0$, la regresión Ridge se convertiría en una estimación de mínimos cuadros, empero, si por el contrario $\lambda \to \infty$ causaría que el impacto de contracción crezca y la estimación de coeficientes se aproxime a cero. A diferencia del método de mínimo cuadrados, el cual genera solamente un conjunto de coeficientes estimados, la regresión Ridge produciría un conjunto diferente de coeficientes estimados, $\hat{\beta}_\lambda^R$, para cada valor λ.

Cabe mencionar que la penalización de contracción afecta a los estimadores β_j exceptuando a β_0, pues lo que se desea con esta regresión es contraer la asociación de cada variable de respuesta, sin embargo, no se quiere contraer al interceptor el cual es simplemente el valor medio de respuesta cuando $x_{i1} = ... = x_{ip} = 0$. Si suponemos que las variables han sido centradas para que tengan media cero previo a ser procesadas por la regresión Ridge, entonces el interceptor estimado tomaría la forma de $\beta_0 = \bar{y} = \sum_{i=1}^{n} y_i/n$.

La ventaja al usar la regresión Ridge sobre el método de mínimos cuadrados se encuentra en el intercambio del bias o sesgo de varianza. Conforme λ crece, la flexibilidad del ajuste de la regresión Ridge se reduce, llevandolo a una reducción de varianza pero a un incremento de bias. Por otro lado, si $\lambda = 0$ entonces los estimadores de mínimos cuadrados tendrían gran varianza pero no contarían con bias.

En general, en situaciones donde la relación entre la variable de respuesta y la variable de predicción sea cercana a una lineal, los estimadores cuadrados tendrían un bias pequeño pero una gran varianza. Esto implicaría que pequeños cambios en los datos de entrenamiento, tendrían un gran efecto en los coeficientes de mínimos cuadrados estimados. En particular cuando el número de variables p sea casi tan grande como el número de observaciones n; y en caso de que $p > n$, entonces los estimadores mínimos cuadrados no tendrían una solución única, mientras que la regresión Ridge puede tener un mejor comportamiento al intercambiar un ligero incremento en el bias por un gran decremento en la varianza. Por

lo tanto, la regresión Ridge funciona mejor en situaciones donde el estimador de mínimo cuadrados tiene una gran varianza.

1.5.3.2. Regresión Lasso

La regresión Ridge tiene una gran desventaja, independientemente del subconjunto de variables que se seleccionen para modelar, la regresión Ridge incluiría todos los p predictores en el modelo final. La penalización $\lambda \sum \beta_j^2$ enviaría a todos los coeficientes hacia cero, pero sin establecerlos en el cero, salvo que $\lambda = \infty$. Esto puede no ser problemático para la precisión de la predicción, sin embargo, puede llegar a ser un reto en la interpretación del modelo cuando el número de variables p sea muy grande.

La regresión Lasso puede ayudar a sortear esta desventaja. Los coeficientes $\hat{\beta}_\lambda^R$, minimiza la cantidad

$$\sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \;=\; \text{SCR} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.57}$$

Con la expresión anterior se pude ver que la gran similitud entre la regresión Ridge y la Lasso, exceptuando las penalizaciones de cada una de las expresión, es decir, el término cuadrado de la regresión Ridge β_j^2 y el término $|\beta_j|$ de la regresión Lasso.

Al igual que la regresión Ridge, la regresión Lasso fuerza a que algunos de sus coeficientes sean cero gracias a su término de penalización, siempre y cuando λ sea lo suficientemente grande. Sin embargo, el seleccionar a λ puede ser complicado.

1.5.3.3. Comparación entre la regresión Ridge y Lasso

A diferencia de la regresión Ridge, la regresión Lasso produce un resultado más sencillo de interpretar pues genera solo un conjunto de estimadores. Por otro lado, tenemos que la varianza generada por la regresión Ridge es ligeramente menor que la varianza de la regresión Lasso, por lo tanto, la mínima SCR de la regresión Ridge es menor que la Lasso.

Sin embargo, la regresión Lasso asume que un cierto número de coeficientes sea iguales a cero, lo que lleva a perder información y que en términos predictivos sea menos precisa que la regresión Ridge. Por otro lado, en términos de bias, varianza y SCR, la regresión Ridge es menos precisa que la regresión Lasso.

En general, se espera que la regresión Lasso supere el comportamiento de la regresión Ridge cuando se tienen un número pequeño de predictores contra un número substancial de

coeficientes; y los predictores restantes tengan coeficientes que sean pequeños o iguales a cero. La regresión Ridge se desempeñará mejor cuando la respuesta se encuentre en función de múltiples predictores, cuyos coeficientes sean muy cercanos en tamaño. Empero, el número de predictores relacionados a la respuesta nunca es conocida *a priori* del conjunto de datos. El método de *cross-validation* puede ser usado para poder determinar cuál aproximación puede resultar mejor para un cierto conjunto de datos.

Además, al igual que la regresión Ridge cuando el último estimador cuadrado tiene una varianza demasiado grande, la solución Lasso nos puede indicar una reducción en la varianza a cambio de incrementar en menor medida el bias, y por ende, generar una predicción más precisa. A diferencia de la regresión Ridge, Lasso mejora la selección de variables y por lo tanto, el modelo resultante tiene una interpretación más sencilla.

Existen varios algoritmos para ajustar de manera eficiente ambos modelos, pues los coeficientes pueden ser calculados con la misma cantidad de esfuerzo que un mínimo cuadrado.

1.5.3.4. Interpretación bayesiana de la regresión Lasso y Ridge

Desde el punto de vista Bayesiano, se asume una distribución *a priori* para el vector β. Supóngase que se tiene el siguiente modelo lineal,

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_p\beta_p + \epsilon \tag{1.58}$$

y supóngase que los errores son normales e independientes. Aún más, supóngase que $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$, para alguna función de densidad g. La regresión Ridge y Lasso se derivan de los dos casos especiales de g.

- Si g es una distribución Gausiana con media cero y desviación estándar en función a λ, entonces β *a posterior*, el cual es el valor más similar a β dado los datos, está dada por la solución de la regresión Ridge. De hecho, la regresión Ridge es la solución de la media posterior.

- Si g es una distribución exponencial doble (Laplace) con media cero y parámetro escalar en función de λ, entonces se sigue que la posterior de β es la solución Lasso. Sin embargo, a diferencia del primer caso, la solución no es la posterior de la media.

1.6. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (MSV) es un enfoque de clasificación desarrollado en la década de los 90's y ha crecido en popularidad desde entonces.

Las MSV es una generalización de un clasificador simple e intuitivo llamado clasificador de margen máximo, sin embargo, este clasificador no puede ser aplicado a la mayoría de los conjuntos de datos, pues requiere que las clases puedan ser separadas por un perímetro lineal.

1.6.1. Clasificador de margen máximo

1.6.1.1. Hiperplanos

En un espacio de dimensión p, un hiperplano se define como un subespacio plano de dimensión $p-1$. Por ejemplo, en dos dimensiones, un hiperplano es un subespacio plano de una dimensión, en otras palabras, una línea.

La definición matemática de hiperplano, en dos dimensiones, es la siguiente

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{1.59}$$

para los parámetros β_0, β_1 y β_2. Cuando se dice que (1.59) "define" el hiperplano, se refiere que cualquier $X = (X_1, X_2)^T$ para el cual (1.59) que se cumpla, es un punto en el hiperplano.

La ecuación (1.59) puede ser fácilmente generalizada para p dimensiones y queda expresado como

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \tag{1.60}$$

Al igual que para dimensión 2, si un punto $X = (X_1, X_2, ..., X_p)^T$ es un espacio de dimensión p y por ende, un vector de longitud p, satisface (1.60) y por lo tanto dicho punto se encuentra en el hiperplano. En caso de que no sea así, significaría que dicho punto x se ha de encontrar en alguno de los dos lados del hiperplano, es decir,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p > 0 \tag{1.61}$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p < 0 \tag{1.62}$$

Es por ello que se puede pensar en el hiperplano como una división del espacio de dimensión p en dos partes.

1.6.1.2. Separación de hiperplanos como método de clasificación

Supóngase que se tiene una matriz X de dimensión n x p que consiste en n observaciones de entrenamiento en un espacio de dimensión p.

$$\begin{matrix} x_{11} & x_{21} & ... & x_{n1} \\ x_{12} & x_{22} & ... & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & ... & x_{np} \end{matrix} \tag{1.63}$$

y estas observaciones caen en dos clases, que son $y_1, ..., y_n \in \{-1, 1\}$ donde -1 representa una clase, mientras que 1 representa la otra clase. También se puede tener una observación de prueba, un vector p de características observadas $x^* = (x_1^*, ..., x_p^*)^T$. El objetivo es desarrollar un clasificador basado en los datos de entrenamiento que clasificaran correctamente la observación de prueba usando sus mediciones de características.

Para la separación por medio de hiperplanos, se pueden etiquetar las observaciones de una clase como $y_i = 1$ y las segundas como $y_i = -1$. Posteriormente un el hiperplano separador debería de tener las características

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} > 0 \quad \text{si} \quad y_i = 1 \tag{1.64}$$

y

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} < 0 \quad \text{si} \quad y_i = -1 \tag{1.65}$$

Equivalentemente, un hiperplano separador tiene la propiedad de

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) > 0 \quad \forall\, i = 1, ..., n \tag{1.66}$$

Si un hiperplano separador existe, se puede usar para construir un clasificador natural. Una observación de prueba es asignada a una clase dependiendo de cuál de los dos lados del hiperplano sea asignada.

Se clasifica la observación de prueba X^* basándose en el signo que $f(x^*) = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + ... + \beta_p X_p^*$. Si $f(x^*)$ es positiva, entonces se ha de asignar a la clase 1, en caso contrario, es decir, que $f(x^*)$ sea negativo, entonces se asigna a la clase -1. También se hace uso de la magnitud de $f(x^*)$, si éste esta lejos del cero, significa que x^* se encuentra lejos del hiperplano y se puede estar seguro sobre la clasificación asignada a x^*. Por el contrario, si $f(x^*)$ se encuentra cercana al cero, implica que x^* se encuentra cerca del hiperplano, y por lo tanto no es tan seguro afirmar la clase a la cual pertenece x^*.

1.6.1.3. El clasificador de margen maximal

Si la información puede ser separada de manera perfecta al usar un hiperplano, entonces existiría una cantidad infinita de hiperplanos tal que hagan lo mismo. Esto se debe a que un hiperplano dado, puede ser movido ligeramente hacia "arriba" o bien hacia "abajo" o bien,

puede ser rotado sin la necesidad de entrar en contacto con alguna de las observaciones que lo rodean.

El hiperplano de margen maximal (también conocido como el hiperplano óptimo de separación), es aquel que se encuentra lo más lejos posible de las observaciones de entrenamiento. Se puede calcular la distancia perpendicular de cada uno de los puntos de entrenamiento, hacia un hiperplano dado, aquel tal que la distancia entre las observaciones y el hiperplano sea la mínima distancia posible. A esto se le conoce como el *marginal*.

El máximo hiperplano marginal es el hiperplano para cuyo margen es mayor, es decir, es el hiperplano que tiene la mayor distancia mínima con las observaciones. Se puede entonces, clasificar una observación de prueba basándose en el lado en el cual se encuentra del hiperplano de margen maximal. A esto se le conoce como el clasificador de margen maximal. Se espera que un clasificador que tenga un gran margen en los datos de entrenamiento, también lo tenga en los datos de prueba, y por lo tanto, pueda clasificar las observaciones de prueba correctamente. A pesar de que el clasificador de margen máximo clasifica satisfactoriamente la mayoría de las veces, también puede caer en sobreajuste cuando p es grande.

Si $\beta_0, \beta_1, ..., \beta_p$ son los coeficientes del hiperplano de margen máximo, entonces el clasificador de margen máximo clasifica las observaciones de prueba x^* basándose en el signo resultante de $f(x^*) = \beta_0 + \beta_1 x_1^* + ... + \beta_p x_p^*$.

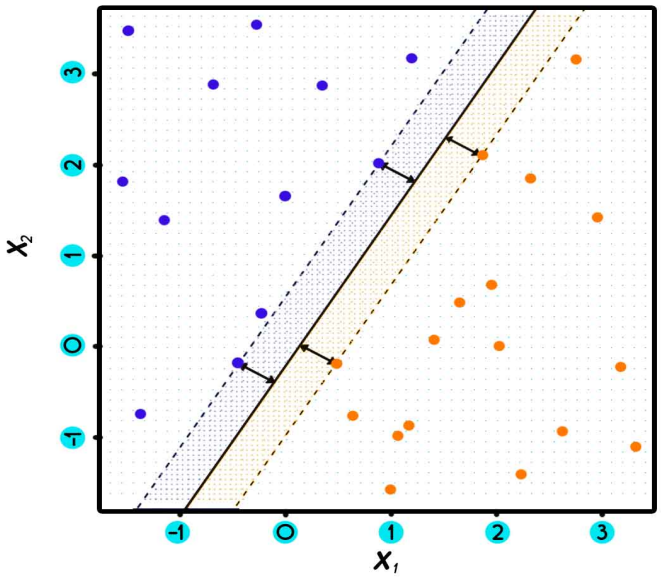


Figura 1.1: *Imagen con dos clases de observaciones, mostradas en amarillo y morado. El margen es la distancia entre las líneas punteadas y la línea sólida. Los dos puntos morados y los dos amarillos en las líneas punteadas son los vectores de soporte. La decisión de clasificación se basa en este hiperplano.*

Se pude ver que tres observaciones de entrenamiento son equidistantes del hiperplano de margen máximo y se encuentran a lo largo de las líneas punteadas, las cuales indican el ancho del margen. Estas tres observaciones son conocidas como soportes vectoriales, dado que estos vectores se encuentran en un espacio de dimensión p y "soportan" el hiperplano de margen máximo tal que, si estos puntos se movieran ligeramente, también lo haría el hiperplano de margen máximo. El hiperplano de margen máximo depende directamente en los soportes vectoriales, pero no así en otras observaciones, es decir el movimiento de cualquier otra observación no debería afectar la separación del hiperplano, siempre y cuando el movimiento de dicha observación no cruce el conjunto de límites del margen.

1.6.1.4. Construcción del clasificador de margen máximo

La construcción del hiperplano de margen máximo basado en un conjunto de n observaciones de entrenamiento $x_1, ..., x_n \in \mathbb{R}^p$ y asociado a clases etiquetadas como $y_1, ..., y_n \in \{-1, 1\}$ puede verse como un problema de optimización, mostrado a continuación

$$\underset{\beta_0,\beta_1,...\beta_p,M}{\text{maximizar}} \; M \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1 \tag{1.67}$$

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) \geq M \ \forall \ i = 1, ..., n \tag{1.68}$$

La restricción (1.68) garantiza que cada observación se encuentre en el lado correcto del hiperplano dado por M positiva, de hecho, simplificando un poco esto, para que cada observación se encuentre realmente en el lado correcto del hiperplano, se necesitaría cumplir con la restricción que $y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) > 0$ por lo que la restricción (1.68) requiere que cada observación se encuentre en el lado correcto del hiperplano, con algún margen para que M sea positivo.

La segunda cosa a notar es que (1.67) no es propiamente una restricción del hiperplano, pues $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} = 0$ define el hiperplano, por lo tanto también lo hace $k(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) = 0$ para cualquier $k \neq 0$. Sin embargo, ayuda a (1.68), se puede demostrar que la restricción de distancia perpendicular de la i-ésima observación al hiperplano esta dada por

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) \tag{1.69}$$

Por lo tanto, la restricción (1.67) y (1.68) aseguran que cada observación se encuentre en el lado correcto y al menos a una distancia M del hiperplano.

1.6.1.5. Caso no separable

En múltiples casos, la separación a través de hiperplanos no existe, por lo que no existe el clasificador máximo marginal. En este caso el problema de optimización visto anteriormente no tiene solución cuando $M > 0$, pues no se puede separar de manera exacta el conjunto de datos en dos clases. Empero, se puede extender el concepto de separación por hiperplanos de tal manera que se desarrolle un hiperplano que "al menos" logre separar las clases, usando lo que se conoce como "margen suave". La generalización del clasificador de margen máximo al caso no separable es conocido como *clasificador de soporte vectorial*.

1.6.2. Clasificador de soporte vectorial

1.6.2.1. Visión general del clasificador de soporte vectorial

Incluso cuando el hiperplano separador no existe, existen instancias en donde un clasificador basado en hiperplanos separadores puede no ser deseable. Un clasificador basado en un hiperplano de separación necesariamente tendría que clasificar perfectamente todos las observaciones de entrenamiento, lo que puede llevar a tener una gran sensibilidad en observaciones individuales. Esto es un problema, pues como se vio anteriormente la distancia entre una observación y el hiperplano puede verse como una medida de confianza de que la observación fue correctamente clasificada. Aún más, el hecho de que el hiperplano de margen máximo es extremadamente sensible a cambios en una sola observación sugiere

que puede tener sobreajustes a los datos.

Por ello, se deberá de considerar un clasificador basado en un hiperplano que no separe de manera perfecta dos clases, para así obtener

1. Un clasificador robusto ante observaciones individuales
2. Un mejor clasificador robusto ante observaciones de entrenamiento

Es decir, podría ser mejor el clasificar que erre algunas observaciones de entrenamiento con la finalidad de hacer un mejor trabajo al clasificar el resto de las observaciones.

El clasificador de soporte vectorial, también conocido como clasificador de margen suave, hace exactamente esto. En lugar que buscar el mayor margen posible de tal manera que las observaciones no solamente caigan en el lado correcto del hiperplano, sino también caigan en el lado correcto del margen, lo cual permite que algunas observaciones se encuentren en lado incorrecto del margen o inclusive en el lado incorrecto del hiperplano.

Las observaciones en el lado incorrecto corresponden a observaciones de entrenamiento que han sido mal clasificadas por el clasificador de soporte vectorial.

1.6.2.2. Detalles del clasificador de soporte vectorial

El clasificador de soporte vectorial clasifica una observación de prueba dependiendo de en que lado del plano recaiga. El hiperplano es elegido para correctamente separar las observaciones de entrenamiento en dos clases, pero pueden existir algunas observaciones que sean mal clasificadas. Esta es la solución a un problema de optimización

$$\underset{\beta_0,\beta_1,...\beta_p,\epsilon_1,...,\epsilon_n,M}{\text{maximizar}} \; M \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1 \tag{1.70}$$

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) \geq \; M(1 - \epsilon_i) \tag{1.71}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C \tag{1.72}$$

donde C es un parámetro de ajuste no negativo. Al igual que en (1.68), M es el ancho del margen, y se busca que esta cantidad sea tan grande como sea posible. En (1.71), $\epsilon_1, ..., \epsilon_n$ son variables de holgura que permitirán a las observaciones individuales estar en el lado equivocado del margen o incluso del hiperplano. Se clasificaría la observación de prueba basándose en el signo que la función $f(x^*) = \beta_0 + \beta_1 x_1^* + ... + \beta_p x_p^*$.

En primer lugar, las variables de holgura ϵ_i indican donde se localiza la i-ésima observación en relación al hiperplano y al margen. Si $\epsilon_i = 0$, significa que la i-ésima observación se encuentra en el lado correcto del margen. Si por el contrario $\epsilon_i > 0$, entonces la i-ésima observación se encuentra en el lado equivocado del margen e indicaría que dicha observación habría violado el margen. Por último, si $\epsilon > 1$ indicaría que la i-ésima observación se encuentra en el lado incorrecto del hiperplano.

En relación a C, esta sería el límite de la suma que los ϵ_i's puedan tener, por lo que determina el número y la severidad de las violaciones que se pueden tolerar al margen y al hiperplano. Se puede pensar en C como un presupuesto para la cantidad de violaciones que pueden tener las n observaciones sobre el margen. En el caso de que $C = 0$, implica que no existe dicho presupuesto y debería de ser el caso en donde $\epsilon_1 = ... = \epsilon_n = 0$, para lo que nos llevaría al primer caso de optimización del hiperplano de margen máximo. Para $C > 0$, las observaciones que se encuentren en el lado incorrecto del hiperplano, no podrían ser mayores a C pues $\epsilon_i > 1$ y (1.72) requiere que $\sum_{i=1}^{n} \epsilon_i \leq C$. Conforme C incremente, el modelo se volvería más tolerante de violaciones al margen.

En la práctica, C es elegido a través de validación cruzada. C controla la compensación del sesgo de varianza de la técnica de aprendizaje estadístico. Cuando C es pequeña, se buscan márgenes estrechos que raramente sean cruzados, lo que implicaría un clasificador que se encuentra muy ajustado a los datos por lo que tendría un menor sesgo pero una mayor varianza. Por otra parte, cuando C sea grande, el margen sería más ancho, lo que permitiría mayores violaciones en él. El ajuste sobre los datos sería menor produciendo un mayor sesgo a cambio de una menor varianza.

El problema de optimización (1.70 - 1.72) posee la propiedad de que únicamente las observaciones que se encuentren en el margen o que violen el margen podrían afectar el hiperplano, y por lo tanto, el clasificador a construir. Una observación que se encuentre estrictamente en el lado correcto del margen no afectaría al clasificador de soporte vectorial. Observaciones que caen directamente en el margen o bien dentro del margen de su clase, sería conocidos como soporte vectorial.

El hecho de que solo los soportes vectoriales afectan al clasificador, mantiene la afirmación anterior, donde C controla la compensación del sesgo de varianza del clasificador de soporte vectorial. Cuando el parámetro de ajuste C es grande, entonces el margen se ampliaría y múltiples observaciones se involucrarían en la determinación del hiperplano.

Una ventaja que presenta la regla de decisión del clasificador de soporte vectorial, es que dependará únicamente en un conjunto potencialmente pequeño de observaciones de entrenamiento, lo que significa que es considerablemente robusto al comportamiento de las observaciones que se encuentran lejos del hiperplano. Esto difiere de otros modelos de

clasificación que necesitan de todos los datos para poder emitir un juicio de clasificación.

1.6.3. Máquinas de soporte vectorial

1.6.3.1. Clasificación con límites de decisión no lineales

El clasificador de soporte vectorial es un enfoque natural para la clasificación de dos clases, si el límite entre ambas clases es lineal. Sin embargo, en la práctica se tienen conjuntos que no necesariamente tienen límites de clase lineal.

En regresión lineal se considera aumentar el espacio de características usando funciones de los predictores, tales como términos cuadráticos o cúbicos de tal forma que se pueda abordar la no linealidad. En el caso del clasificador de vectores de soporte, se puede abordar este problema de una forma similar, al incrementar las características del espacio usando funciones polinomiales, cuadráticas, cubicas o de algún orden superior, de los predictores en lugar de ajustar clasificadores de soporte vectorial usando p características.

$$X_1, X_2, ..., X_p \tag{1.73}$$

En lugar de eso, se puede ajustar un clasificador de soporte vectorial usando 2p características

$$X_1,\ X_1^2,\ X_2,\ X_2^2, ..., X_p,\ X_p^2 \tag{1.74}$$

Entonces, el problema de optimización (1.70 - 1.72) se convierte en

$$\underset{\beta_0,\beta_{11},\beta_{12},...\beta_{p1},\beta_{p2},\epsilon_1,...,\epsilon_n,M}{\text{maximizar}} M \text{ sujeto a } \quad y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} \ + \ \sum_{j=1}^{p} \beta_{j2} x_{ij}^2 \right) \geq M(1-\epsilon_i)$$
$$\sum_{i=1}^{n} \epsilon_i \leq C, \quad \epsilon_i \geq 0, \quad \sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1 \tag{1.75}$$

Al aumentar las características del espacio, la decisión del límite resultante de (1.75) es de hecho lineal. Pero en el espacio original, el límite de decisión es de la forma $q(x) = 0$, donde q es un polinomio cuadrático y cuya solución generalmente es no lineal. Se podría incrementar las características del espacio con un polinomio de orden mayor o bien con la interacción de los términos $X_j X_{j'}$ para $j \neq j'$. Empero, esta no es la única forma de hacer crecer el espacio de características, pues existen otras funciones de predictores que pueden cumplir con dicha tarea sin la necesidad de caer en procedimientos computacionalmente complejos.

1.6.3.2. Uso del kernel en las MSV (Truco del kernel)

Las máquinas de soporte vectorial (MSV) es una extensión del clasificador de soporte vectorial que se obtiene como resultado de incrementar el espacio de características usando kernels. El enfoque a través de kernels es una forma eficiente computacionalmente hablando.

La clasificación de vectores de soporte, en esencia, busca clasificar los datos que se tienen en varias clases objetivo, siempre y cuando estas puedan ser separadas por alguna línea o algún límite. De esta manera se pueden clasificar los datos dependiendo del lado en el cuál se encuentren los datos. En la teoría, es fácil separar estos grupos de información, sin embargo, en la práctica los datos están lejos de ser separables linealmente y es por ello que es necesario transformar los datos en un espacio de dimensiones más grande de tal manera de que se pueda ajustar un clasificador de vectores de soporte.

A esto se le conoce como el truco del kernel, es decir, en el caso de que los datos no puedan ser separados linealmente en el espacio original, entonces habrá que aplicarse alguna transformación (por medio del kernel) de tal manera de que se puedan mapear los datos en un espacio de características de mayor dimensión. El objetivo final de esta transformación es que las clases sean linealmente separables en este nuevo espacio.

Sin embargo, se puede ver que la solución para el problema del clasificador de soporte vectorial (1.70 - 1.72) incluye sólo los productos internos de las observaciones. El producto interno de dos r-vectores a y b se define como $\langle a, b\rangle = \sum_{i=1}^{r} a_i b_i$. Por lo que, el producto interno de dos observaciones $X_i, X_{i'}$ está dado por

$$\langle X_i, X_{i^*}\rangle = \sum_{i=1}^{r} X_{ij}, X_{i^*j} \tag{1.76}$$

Puede mostrarse que

1. El clasificador lineal de soporte vectorial puede ser representado como

$$f(x) \quad = \quad \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i\rangle \tag{1.77}$$

   donde existen n parámetros $\alpha_i,\ i = 1, ..., n$, uno por cada observación de entrenamiento.

2. Para estimar los parámetros $\alpha_i, ..., \alpha_n$ y β_0 todo lo que se necesita es $\binom{n}{2}$ productos internos $\langle x_i, x_i*\rangle$ entre todos los pares de observaciones de entrenamiento.

Nótese que en (1.77), para evaluar la función $f(x)$, se necesita calcular el producto interno entre el punto nuevo x y cada uno de los puntos de entrenamiento x_i. Sin embargo, resulta que α no es cero solo para la solución del soporte vectorial. Entonces, si S es la colección de índices de estos puntos de soporte, se puede reescribir la solución de la forma (1.77), como

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle \tag{1.78}$$

Supóngase que en cada momento que el producto interno (1.76) aparece en la representación (1.77), o en el cálculo de la solución del clasificador del soporte vectorial, se reemplace con la generalización del producto interno de la forma

$$K(x_i, x_{i^*}) \tag{1.79}$$

donde K es alguna función referente al *kernel*. El kernel es la función que cuantifica la similaridad entre dos observaciones. Por ejemplo, se puede tomar el caso donde

$$K(x_i, x_{i^*}) = \sum_{j=1}^{p} x_{ij} x_{i^* j} \tag{1.80}$$

donde ha de regresar el clasificador de soporte vectorial. A la ecuación (1.80) se le conoce como kernel lineal pues el clasificador de soporte vectorial es lineal en sus características. El kernel lineal en esencia cuantifica la similaridad de un par de observaciones usando la correlación estándar de Pearson, aunque puede no ser la única. Por ejemplo en el caso de (1.79) se puede sustituir cada sección de $\sum_{j=1}^{p} x_{ij} x_{i^* j}$ con la cantidad

$$K(x_i, x_{i^*}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i^* j})^d \tag{1.81}$$

A esta ecuación se le conoce como *kernel polinomial* de grado d, donde d es un entero positivo. Usando dicho kernel con $d > 1$, en lugar de un kernel lineal estándar (1.80), en el algoritmo del clasificador de soporte vectorial se permite un límite más flexible. Esencialmente, equivale a ajustar un clasificador de soporte vectorial en un espacio de gran dimensión usando un polinomio de grado d, en lugar de usar el espacio original. Cuando el clasificador de soporte vectorial se combina con un kernel no lineal, como en (1.81), el resultado de este clasificador es conocido como Máquina de Soporte Vectorial (MSV). Nótese que en este caso, no lineal, la función adopta la forma de

$$f(X) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{1.82}$$

Cuando $d = 1$, la MSV se reduce a un clasificador de soporte vectorial.

Otra posible elección es el *kernel radial* la cual toma la siguiente forma:

$$K(x_i, x_{i^*}) = \exp\left(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i^*j})^2\right) \tag{1.83}$$

donde γ es positivo constante.

Un kernel radial es una observación de prueba dada $x^* = (x_1^*, ..., x_p^*)^T$, que se encuentra lejos de las observaciones de entrenamiento x_i en términos de distancia euclidiana, entonces $\sum_{j=1}^{p}(x_j^* - x_{ij})^2$ sería mayor, y por lo tanto $K(x_i, x_{i^*}) = \exp\left(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i^*j})^2\right)$ sería muy pequeña. Esto significa que en (1.82) x_i no tendría participación en $f(x^*)$. Recordar que la clase predicha para las observaciones de prueba x^* están basadas en el signo resultante de $f(x^*)$. En otras palabras, las observaciones de entrenamiento se encuentren alejadas de x^* no tendrían ninguna repercusión en la predicción de la clase de x^*. Por lo que se puede decir que el kernel radial tiene un fuerte comportamiento local, en el sentido de que solo las observaciones de entrenamiento que se encuentren cerca tienen efecto en la clase que tendría la observación de prueba.

Algunos otros kernels que llegan a ser populares son los siguientes:

- Laplace RBF, es un kernel de propósito general, usada cuando no se tiene información previa a cerca de los datos

$$k(x, y) = \exp\left(-\frac{||x - y||}{\sigma}\right) \tag{1.84}$$

- Tangente hiperbólico, usado normalmente en redes neuronales

$$k(x_i, x_j) = \tanh(kx_i * x_j + c) \tag{1.85}$$

  para alguna $k > 0$ y $c < 0$

- Sigmoideo, se puede usar como proxy para redes neuronales.

$$k(x, y) = \tanh(\alpha x^T y + c) \tag{1.86}$$

La ventaja de usar el kernel en lugar de aumentar las características del espacio es principalmente computacional, pues al usar el kernel, sólo se hace el cálculo $K(x_i, x_{i^*})$ para todos los $\binom{n}{2}$ de pares i, i^*. Esto puede realizarse sin la necesidad de aumentar el espacio de características de manera explícita.

1.6.4. MSV con más de dos clases

El concepto de separar dos hiperplanos es donde se encuentra basada la teoría de MSV y no es fácil pensar bajo este concepto en separar más de dos clases. Sin embargo, existen un par de propuestas para poder extender las MSV en K-clases, dos de las más populares son uno-contra-uno y uno-contra-todos.

1.6.4.1. Clasificación uno contra uno

Supóngase que se quiere clasificar K clases, con $K > 2$ a través de MSV. El método uno-contra-uno, también llamado total de pares, es una construcción $\binom{K}{2}$ de MSV, donde cada una compara un par de clases.

Por ejemplo, una sola MSV compara la k-ésima clase, con etiqueta $+1$ con las k clases etiquetadas con -1. Se clasifica una observación de prueba usando cada una de los $\binom{K}{2}$ clasificadores y se cuenta el número de veces que la observación de prueba es asignada a cada una de las K clases. La última clasificación se realiza al asignar la observación de prueba a la clase cuya frecuencia haya sido asignada más veces en los pares de clasificación $\binom{K}{2}$.

1.6.4.2. Clasificación uno contra todos

Uno-contra-todos es un procedimiento alternativo para aplicar MSV en el caso de que $K > 2$. Se ajusta K MSV, siempre comparando cada una de las K clases contra las $K-1$ clases restantes. Sean $\beta_{0k}, ..., \beta_{pk}$ los parámetros que resultan de ajustar un MSV comparando la k-ésima clase, etiquetada como $+1$, con respecto de las otras, etiquetadas como -1.

Sea x^* una observación de prueba, se asigna la observación a la clase a la cual $\beta_{0k} + \beta_{1k}x_1^* + ... + \beta_{pk}x_1^*$ es mayor, lo que equivale a un mayor nivel de confianza de que la observación de prueba pertenece a la k-esima clase en lugar de pertenecer a cualquier otra clase.

1.6.5. Relación con la regresión logística

Desde la creación de las MSV en la década de los 90's se han creado más conexiones entre las MSV y otros métodos de estadística clásica. Lo que ha llevado a poder reescribir los criterios (1.70 - 1.72) al ajustar un clasificador de soporte vectorial $f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ como

$$\underset{\beta_0,\beta_1,...\beta_p}{\text{maximizar}} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i f(x)] + \lambda \sum_{j_1}^{p} \beta_j^2 \right\} \tag{1.87}$$

donde λ es un parámetro de ajuste no negativo, cuando éste se hace grande, el clasificador resultante tendría como características $\beta_1, ..., \beta_p$ pequeños, el tolerar más violaciones del margen y existiría una varianza pequeña con un sesgo grande. Por el otro lado, cuando λ sea pequeño, entonces el clasificador tendría pocas violaciones al margen, una gran varianza pero un menor sesgo. Por lo tanto, un valor pequeño para λ en (1.87) equivale al valor C en (1.72). Nótese que $\lambda \sum_{j=1}^{p} \beta_j^2$ en términos de (1.87) es el término de penalización *Ridge* y juega un rol similar en el control del sesgo y la varianza para el clasificador de soporte vectorial.

Ahora (1.87) toma la forma de "pérdida y penalización"

$$\underset{\beta_0,\beta_1,...\beta_p}{\text{minimizar}} \{L(X, y, \beta) + \lambda P(B)\} \tag{1.88}$$

En (1.88), $L(X, y, \beta)$ es una función de perdida que cuantifica el grado en el cual el modelo, parametrizado con β, ajusta a los datos (X, y) y $P(B)$ es una función de penalización en el parámetro del vector β cuyo efecto es controlado a través de un parámetro no negativo de ajuste λ.

Ejemplo 1.6.1.

La regresión Ridge y Lasso toman la siguiente forma

$$L(X, y, \beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{1.89}$$

y con $P(B) = \sum_{j=1}^{p} \beta_j^2$ para la regresión Ridge y $P(B) = \sum_{j=1}^{p} |\beta_j|$ para Lasso. En el caso de (1.87) la función de pérdida toma la forma de

$$L(X, y, \beta) = \sum_{i=1}^{n} \max[0,\ 1\ -\ y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})] \tag{1.90}$$

Esta forma es conocida como función eje de pérdida o *hinge loss* en inglés. Sin embargo, resulta que la función de eje de pérdida se encuentra relacionada con la función de pérdida de la regresión logística.

∇

A diferencia de las MSV que únicamente se ven afectadas por los puntos cercanos al clasificador y no a aquellos que se encuentran alejados de éste, la función de pérdida de la regresión logística no es exactamente cero en ningún punto. Pero llega a ser muy pequeña para observaciones que se encuentran lejos del límite de decisión.

Debido a las similaridad entre estas funciones de pérdida, ambos métodos (clasificador de soporte vectorial y regresión logística) normalmente obtienen resultados similares. Cuando las clases están bien separadas, las MSV tienen un mejor comportamiento que las regresiones logísticas.

Pese a que el clasificador de soporte vectorial está estrechamente relacionado con la regresión logística y otros métodos estadísticos, las MSV son las únicas en usar kernels para aumentar el espacio de características para establecer límites de clase no lineales.

Existe una extensión de MSV para regresión, es decir, para características cuantitativas en lugar de cualitativas, llamada Regresión de Soporte Vectorial. Este método busca coeficientes en lugar de minimizar el coeficiente de pérdida; donde solo los mayores residuales en valor absoluto contribuyen a la función de pérdida, en lugar de alguna constante positiva. Esta es una extensión del límite utilizado en los clasificadores de soporte vectorial a la configuración de regresión.

1.7. Bosques aleatorios

1.7.1. Introducción a técnicas basadas en árboles

Supóngase que se tiene un problema de clasificación. Cada variable y_i denota cierta clase. Supóngase que hay K clases y que están etiquetadas por los números $1, 2, \ldots, K$, respectivamente. Como estas son clases, no hay una relación de orden entre las clases.

Considérese un árbol de decisión con m hojas, denotadas por $R_1, \ldots, R_m$. El número de observaciones que caen en la hoja R_j se denotará por n_j.

Se debe predecir a qué clase pertenece un vector predictor x_i. Si el vector predictor cae en la región R_j y la función de aprendizaje se denota por $g(\cdot)$, se está intentando determinar $g(x_i)$ si $x_i \in R_j$.

Un árbol de clasificación es adecuado para describir los datos de entrenamiento si el árbol es capaz de asignar la clase correcta a la mayoría de las observaciones.

En cada hoja del árbol se tiene que decidir qué clase asignar. La predicción en cierta hoja corresponde a la clase que más se repite en dicha hoja.

Esta metodología se reduce a minimizar la tasa de *misclassification*, i.e. minimizar la probabilidad de que una observación seleccionada aleatoriamente de esta hoja esté mal clasificada.

En un nodo particular del árbol, la mejor situación ocurre cuando todas las observaciones pertenecen a la misma clase. En esta situación no hay duda cuando se asigna la predicción para esta hoja y todas las observaciones en esta hoja están correctamente clasificadas.

Definición 1.7.1. *(Nodo puro de un árbol)*

Un nodo en el que todas las observaciones pertencen a la misma clase se le conoce como nodo puro.

Supóngase que se construyó un árbol de clasificación con m hojas. La mejor situación en este árbol de clasificación ocurre cuando todas las m hojas del árbol son nodos puros. En este caso, se puede clasificar a todas las observaciones de entrenamiento correctamente, usando las m hojas.

Cuando se construye un árbol de clasificación, se separa cada nodo en dos nodos hijos. Se están buscando splits que generen nodos hijos tan puros como sea posible. Por lo tanto se necesita cuantificar el grado de pureza (ó equivalentemente de impureza) de un nodo.

Supóngase que el grado de impureza del nodo de la hoja R_j se cuantifica mediante ρ_j. Por lo tanto, si la hoja R_j es pura, entonces $\rho_j = 0$. Mientras más grande sea ρ_j, más impura será la hoja R_j.

1.7.2. Medición del grado de impureza de un nodo

Definición 1.7.2. *(Frecuencia de clase)*

Considérese un árbol de clasificación con m hojas. Se define la frecuencia de la clase k en la hoja j, $\hat{p}_{j,k}$, como

$$\hat{p}_{j,k} := \frac{1}{n_j} \sum_{x_i \in R_j} I(y_i = k) \tag{1.91}$$

La frecuencia $\hat{p}_{j,k}$ se puede interpretar como la probabilidad empírica de que una observación sea de la clase k dado que la observación pertenece a la hoja R_j.

Se considerarán tres medidas de la impureza de un nodo:

(i) Misclassification error.

(ii) Índice de Gini.

(iii) Cross-entropy.

Definición 1.7.3. *(Misclassification error de una hoja)*

Se define el misclassification error de una hoja R_j como

$$E_j = 1 - \max_k \hat{p}_{j,k} \tag{1.92}$$

Observación 1.7.1.

Para cualquier $j \in \{1, \ldots, m\}$

$$\hat{p}_{j,1}, \hat{p}_{j,2}, \ldots, \hat{p}_{j,K} \leq \max_k \hat{p}_{j,k} \tag{1.93}$$

equivalentemente

$$1 - \hat{p}_{j,1}, 1 - \hat{p}_{j,2}, \ldots, 1 - \hat{p}_{j,K} \geq E_j \tag{1.94}$$

es decir, la probabilidad de que una observación esté mal clasificada está acotada por el misclassification error. ▽

Si el *misclassification error* es cercano a 0, esto significa que la mayoría de los datos que están en la hoja R_j pertenecen a la misma clase. Esto se debe a que $E_j = 0$ si y sólo si $1 = \max_k \hat{p}_{j,k}$ si y sólo si existe $l \in \{1, \ldots, K\}$ tal que $\hat{p}_{j,l} = 1$. Es decir, $n_j = \sum_{x_i \in R_j} I(y_i = l)$. Por lo tanto todos los y_i's en la hoja R_j son de la clase l, i.e. la hoja es pura.

Entonces, mientras más alto sea E_j, más alta será la impureza de R_j (equivalentemente, mientras más bajo sea E_j, más baja será la pureza de R_j); y mientras más alto sea el grado de impureza, más difícil será predecir la clase correctamente en esa hoja particular.

Misclassification para el caso bidimensional

En particular, para clasificación binaria se pueden hacer algunos cálculos más especificos.

Considérese un nodo R_j de un árbol de clasificación. Supóngase también, que hay dos clases para la variable respuesta: 0 ó 1.

El error de mala clasificación está dado por

$$E_j = 1 - \max \hat{p}_{j,0}, \hat{p}_{j,1} \tag{1.95}$$

Se puede graficar la aplicación $\hat{p}_{j,0} \mapsto E_j$

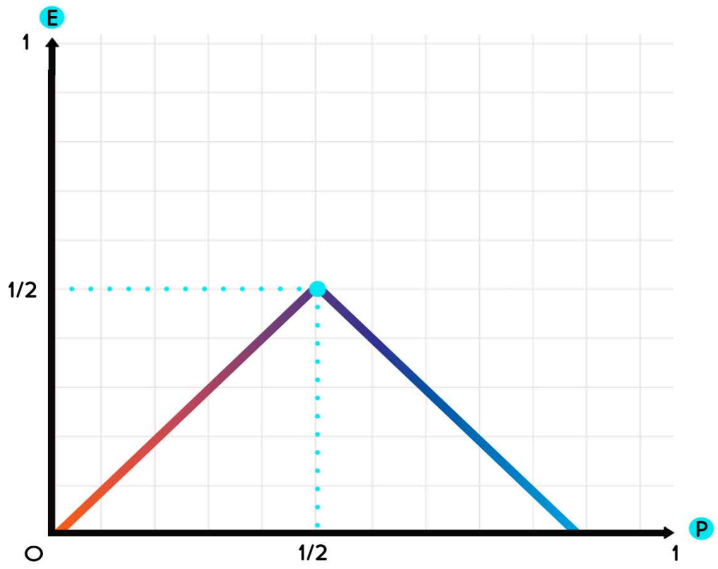


Figura 1.2

Obsérvese que esta función no es diferenciable en $\hat{p}_{j,0} = 1/2$. Para problemas de optimización, esta propiedad no es deseable. Esto motiva la definición del índice de Gini y la cross-entropy.

Definición 1.7.4. *(Indice de Gini de una hoja)*

Se define el índice de Gini para la hoja R_j como

$$G_j := \sum_{k=1}^{K} \hat{p}_{j,k}(1 - \hat{p}_{j,k}) \tag{1.96}$$

Definición 1.7.5. *(Cross-entropy para una hoja)*

Se define la cross-entropy para la hoja R_j como

$$D_j := -\sum_{k=1}^{K} \hat{p}_{j,k} \log(\hat{p}_{j,k}) \tag{1.97}$$

Como en el caso del error de mala clasificació, se pueden hacer algunos cálculos para el caso de clasificación binaria.

El índice de Gini es

$$\begin{aligned} G_j &= \hat{p}_{j,0}(1-\hat{p}_{j,0}) + \hat{p}_{j,1}(1-\hat{p}_{j,1}) \\ &= \hat{p}_{j,0}(1-\hat{p}_{j,0}) + (1-\hat{p}_{j,0})\hat{p}_{j,0} \\ &= 2\hat{p}_{j,0}(1-\hat{p}_{j,0}) \\ &= 2\hat{p}_{j,0} - 2\hat{p}_{j,0}^2 \end{aligned}$$

Obsérvese que la aplicación $\hat{p}_{j,0} \mapsto G_j$ es diferenciable, y de hecho alcanza su máximo en $\hat{p}_{j,0} = 1/2$. Esto quiere decir que si se usa al índice de Gini como medida de impureza, la impureza se maximiza en $\hat{p}_{j,0} = 1/2$, lo que significa que la mitad de las observaciones en la hoja pertenecen a la clase 0 y la otra mitad a la clase 1.

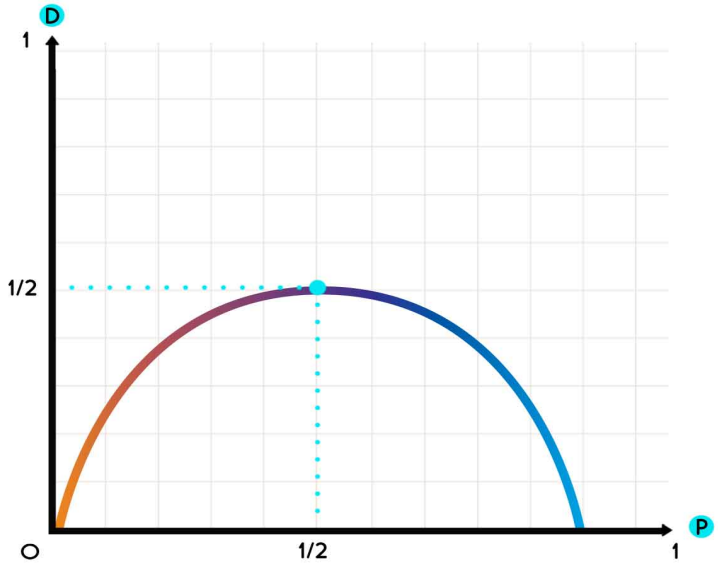


Figura 1.3

Nótese también que si $\hat{p}_{j,0} = 1$, entonces todas las observaciones en la hoja R_j pertenecen a la clase 0. Análogamente, si $\hat{p}_{j,0} = 0$, entonces todas las observaciones en la hoja R_j pertenecen a la clase 1.

En este caso de dos categorías, también la cross-entropy de la hoja R_j es

$$\begin{aligned} D_j &= -[\hat{p}_{j,0}\log(\hat{p}_{j,0}) + \hat{p}_{j,1}\log(\hat{p}_{j,1})] \\ &= -\hat{p}_{j,0}\log(\hat{p}_{j,0}) - (1-\hat{p}_{j,0})\log(1-\hat{p}_{j,0}) \end{aligned}$$

Obsérvese que la aplicación $\hat{p}_{j,0} \mapsto D_j$ es diferenciable y además

$$\begin{aligned} \frac{\partial}{\partial \hat{p}_{j,0}} D_j &= -\hat{p}_{j,0}\frac{1}{\hat{p}_{j,0}} - \log(\hat{p}_{j,0}) - (1-\hat{p}_{j,0})\left(-\frac{1}{(1-\hat{p}_{j,0})}\right) + \log(1-\hat{p}_{j,0}) \\ &= \log\left(\frac{1-\hat{p}_{j,0}}{\hat{p}_{j,0}}\right) \end{aligned}$$

De aquí que $\frac{\partial}{\partial \hat{p}_{j,0}} D_j = 0$ si y sólo si

$$\frac{1-\hat{p}_{j,0}}{\hat{p}_{j,0}} = 1$$

es decir si y sólo si $\hat{p}_{j,0} = 1/2$. Y como

$$\frac{\partial^2}{\partial \hat{p}_{j,0}^2} D_j = -\frac{1}{1-\hat{p}_{j,0}} - \frac{1}{\hat{p}_{j,0}} < 0$$

se tiene que $\hat{p}_{j,0} = 1/2$ es máximo de D_j.

Como en en el caso del índice de Gini, la aplicación $\hat{p}_{j,0} \mapsto D_j$ alcanza su máximo en $\hat{p}_{j,0} = 1/2$. Esto quiere decir que si se usa a la cross-entropy como medida de impureza, la impureza se maximiza en $\hat{p}_{j,0} = 1/2$, lo que significa que la mitad de las observaciones en la hoja pertenecen a la clase 0 y la otra mitad a la clase 1.

Los métodos de árboles se pueden dividir en dos, en aquellos que se aplican para regresión y los que se aplican para clasificación. Esto incluye la estratificación y segmentación en un cierto número de regiones.

De igual manera se hablará sobre los métodos de *bagging, bosques aleatorios* y *boosting* los cuáles se combinarán para llegar en consenso a una única predicción.

1.7.3. Árboles de decisión básicos

Debido a que el conjunto de reglas de división usadas para segmentar el espacio de predicción puede ser representado de manera resumida como un árbol, este tipo enfoque puede ser conocido como método de "Árboles de decisión"‘.

Usualmente el desempeño de estos métodos no se comporta tan bien como lo hacen los métodos bajo el enfoque de aprendizaje supervisado.

1.7.4. Árboles de regresión

Los árboles de decisión pueden aplicarse tanto a problemas de clasificación, como de regresión. Se comenzará a revisar los problemas de regresión y posteriormente se analizará la aplicación para clasificación.

Los árboles de decisión están conformados por un nodo inicial del cuál partirá la primer división. Esta primer división generará dos regiones iniciales conocidas como R_1 y R_2, eventualmente estas regiones se partirán en tantas decisiones como sea necesario en pro de obtener un resultado más exacto, hasta llegar a las regiones finales, también conocidas como "nodos finales" o "hojas de árbol".

Estos árboles de decisión, debido a como se van desarrollando, son generalmente dibujados de arriba hacia abajo, de tal manera que las "hojas del árbol", se encuentran en la parte inferior del mismo. Los puntos que se encuentran a lo largo del árbol, donde se hace la

división de los predictores de espacio son conocidos como nodos internos; y a los segmentos del árbol que salen de estos nodos los conoceremos como "ramas".

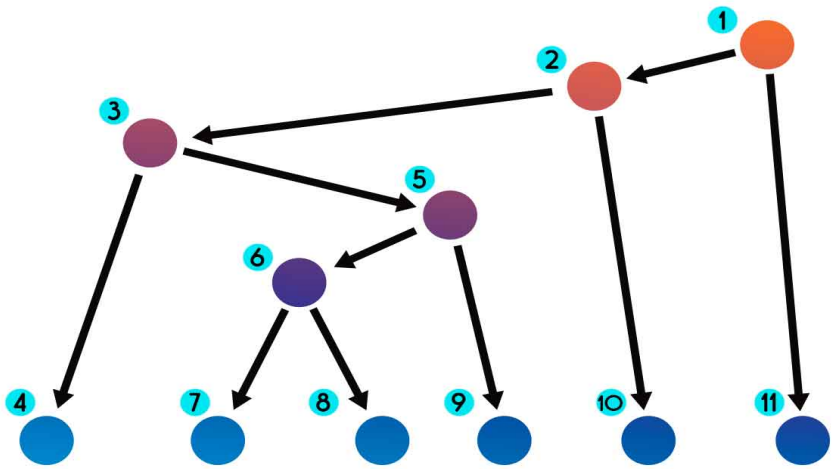


Figura 1.4: *Árbol de decisión*

1.7.4.1. Predicción vía estratificación del espacio de características

El proceso de construir un árbol de regresión se compone de dos pasos básicamente:

1. Se divide el espacio de predicción, el cuál es el conjunto de posibles valores para $X_1, X_2, ..., X_P$ en J regiones distintas y no intersectadas, las cuáles llamaremos $R_1, R_2, ..., R_J$.

2. Parra cualquier observación que caiga dentro de la región R_J, se hará la misma predicción, la cual es simplemente la media de los valores de respuesta de los datos de entrenamiento para R_J.

Supóngase que en el paso 1 se obtuvieron dos regiones, R_1 y R_2, y la media de respuesta de los valores de entrenamiento para la primer región es a, mientras que la media de respuesta de la segunda región es b, $a, b \in \mathbb{R}$. Entonces dada la observación $X = x$, si $x \in R_1$, entonces el valor predicho será de a, mientras que para $x \in R_2$, el valor obtenido será de b.

Para facilitar su representación gráfica, se han de dividir los espacios de predicción en rectángulos o cajas. El objetivo es encontrar las cajas $R_1, ..., R_J$ que minimice la Suma de Residuales Cuadrados (SRC), dada por

$$\sum_{j=1}^{J} \sum_{i \in R_J} (y_i - \hat{y}_{R_J})^2 \tag{1.98}$$

donde $\hat{y}_{R_J}$ es la media de respuesta de las observaciones de entrenamiento dentro de la j-ésima caja. Empero, este cálculo puede ser computacionalmente ineficiente al considerar toda posibilidad de división de los espacios de características dentro de las J cajas. Para evitar este problema, se ha de utilizar el método "glotón" de *División Binaria Recursiva* de "arriba hacia abajo".

Tiene ese nombre ya que comenzará en la parte superior del árbol (en dónde todas las observaciones pertenecerán a una sola región) y posteriormente se dividirán los espacios de predicción, cada una de las divisiones se harán en dos nuevas ramas que bajarán en el diagrama de árbol. El nombre de glotón es debido a que en cada uno de los pasos de la construcción de este árbol, la mejor división es hecho en ese paso en particular, en lugar de buscar una división más adelante y poder generar un mejor árbol en un paso futuro.

A fin de implementar la división binaria recursiva, primero se ha de seleccionar el predictor X_J y los puntos de corte s tal que al dividir el espacio de predicción en regiones $\{X|X_j < s\}$ y $\{X|X_j \geq s\}$ se lleva a la mayor reducción posible de la SRC. De esta manera se consideran todos los predictores $X_1, ...X_p$ y a todos los posibles valores de los puntos de corte s para cada uno de estos predictores; y posteriormente elegir el predictor y el punto de corte tal que el árbol resultante tenga la menor SRC. En otras palabras, para cada j y s, se define el par de medios-planos

$$R_1(j,s) = \{X|X_j < s\} \quad \text{y} \quad R_2(j,s) = \{X|X_j \geq s\} \tag{1.99}$$

y se buscan los valores de j y s tal que se minimiza la ecuación

$$\sum_{i:\ x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\ x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{1.100}$$

donde $\hat{y}_{R_1}$ es la media de respuesta de las observaciones d entrenamiento en $R_1(j,s)$ y $\hat{y}_{R_2}$ es la media de respuesta de las observaciones de entrenamiento de $R_2(j,s)$.

Después, se ha de repetir este proceso eligiendo los mejores predictores y los mejores puntos de corte con el fin de dividir los datos de tal manera de poder minimizar la SRC dentro de cada una de las regiones resultantes. Sin embargo, en esta ocasión a diferencia de dividir todos los espacios de predicción, se ha de dividir en dos la región previamente

identificada. Se ha de continuar de esta manera hasta que cada una de las regiones tenga no más de 4 observaciones.

Una vez que las regiones $R_1, ..., R_J$ han sido creadas, se procede a predecir la respuesta para una observación de prueba dada, usando la media de las observaciones de entrenamiento en la región a la cual la observación de prueba pertenece.

1.7.4.2. Poda del árbol

Un árbol pequeño con pocas divisiones (es decir, pocas regiones $R_1, ..., R_J$) puede tener una menor varianza y una mejor interpretación, aumentando un poco el sesgo. Una posibilidad para el proceso anteriormente descrito es la construcción de un sólo árbol que crezca siempre y cuando la SRC en cada división exceda cierto límite. Esta estrategia resultará en árboles más pequeños, pero puede resultar una estrategia no tan óptima al hacer divisiones sin sentido al inicio del árbol, seguidas por mejores divisiones, lo que llevaría a una mayor reducción en la SRC más tarde.

Por lo tanto una mejor estrategia es hacer crecer un árbol T_0 y posteriormente *podarlo* para así obtener un sub-árbol, el objetivo es seleccionar el árbol con menor tasa de error en las pruebas. Se puede estimar el error de prueba al aplicarle una validación cruzada o el conjunto de validación. Sin embargo, el estimar la validación cruzada para cada uno de los posibles sub-árboles puede llegar a ser bastante complejo.

La poda de enlace débil o La poda de complejidad de costos, da una forma de hacer esto último. En lugar de considerar cada uno de los subárboles posibles, se considera una secuencia de árboles indexados por un parámetro no negativo de ajuste α. Para cada valor de α corresponde un subárbol $T \subset T_0$ tal que

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{1.101}$$

sea tan pequeña como sea posible. En este caso $|T|$ indica el número de nodos terminales del árbol T, R_m es la caja correspondiente al m-ésimo nodo terminal y $\hat{y}_{R_m}$ es el predictor responsable asociado con $\hat{y}_{R_m}$, es decir, la media de las observaciones de entrenamiento en R_m. El parámetro de ajuste α controla la compensación entre el la complejidad del subárbol y su ajuste con los datos de entrenamiento. Cuando $\alpha = 0$, entonces el subárbol T será simplemente igual al árbol padre T_0, pues (1.101) medirá justamente el error de entrenamiento. Empero, conforme α crezca, la cantidad derivada de (1.101) tenderá a ser minimizada para un subárbol pequeño. La ecuación (1.101) es una reminiscencia de la ecuación Lasso (1.57), la cual es similar y usada para controlar la complejidad del modelo lineal.

Como resultado de incrementar α, las ramas del árbol van siendo podadas de una forma anidada y predecible, por lo tanto, se obtiene la secuencia completa de subárboles como función de α de manera sencilla. Se puede seleccionar el valor de α al usar un conjunto de validación o usar validación cruzada. Después se ha de regresar a los datos completos y obtener el subárbol correspondiente a dicho α. Este proceso es descrito más a detalle a continuación.

1.7.4.3. Construcción de un árbol de regresión

1. Usar división binaria recursiva para hacer crecer el árbol con los datos de entrenamiento hasta que el número de observaciones sea el mínimo.

2. Aplicar la poda de complejidad de costos a grandes árboles para así obtener una secuencia de los mejores subárboles, como función de α.

3. Usar el método de validación cruzada de K-pliegues para elegir α. Que es, dividir las observaciones de entrenamiento en K-pliegues. Para cada $1, ..., K$:

   - Repetir los pasos 1 y 2 en todos menos el k-ésimo pliegue de datos de entrenamiento.
   - Evaluar el error de la media cuadrada de predicción de los datos en el excluido k-ésimo pliegue, como función de α.

   Promediar el resultado para cada uno de los valores de α, y tomar α para minimizar el error promedio.

4. Regresar con el subárbol al paso 2 que corresponde a elegir un valor para α.

1.7.5. Árboles de clasificación

El árbol de clasificación es usado para predecir la respuesta cualitativa en lugar de una respuesta cuantitativa. Para este árbol, se predice que cada observación pertenece a la clase más común de ocurrencia de las observaciones de entrenamiento, en la región a la cual pertenece. La interpretación de los resultados relacionados al árbol de clasificación, no solo es interesante en términos de la clase correspondiente predicha, sino también en la proporción que dicha clase con las observaciones de entrenamiento que caen en dicha región.

La construcción de un árbol de clasificación es muy similar a la construcción del árbol de regresión, sin embargo, en el primero la SRC no puede ser usada como criterio para hacer la división binaria. Una alternativa a esta situación, es la tasa de error de clasificación; al asignar una observación de una región dada a la clase más común de ocurrencia de las observaciones de entrenamiento en la región, la tasa de error de clasificación es una fracción

de las observaciones de entrenamiento. Es esa región que no pertenecen a la clase más común, esto puede ser calculado de la siguiente manera:

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{1.102}$$

Para este caso, $\hat{p}_{mk}$ representa la proporción de observaciones de entrenamiento en la m-ésima región que pertenece a la k-ésima clase. Sin embargo, en la práctica, este método no es lo suficientemente preciso para un árbol en crecimiento y es preferible optar por las siguientes dos medidas.

El índice de Gini, definición (7.5), es una medida de total varianza a lo largo de las K clases. El índice de Gini es pequeño si todas las $\hat{p}_{mk}$'s son cercanos a cero o a uno. Por esta razón, el índice de Gini se refiere como una medida de pureza del nodo, un valor pequeño indica que el nodo contiene predominantemente observaciones de una sola clase.

Definición 1.7.6. *(Entropia)*

Se define a la entropia, como alternativa al índice de Gini, como

$$D = -\sum_{k=1}^{K} \hat{p}_{mk}\ log(\hat{p}_{mk}) \tag{1.103}$$

Si $0 \leq \hat{p}_{mk} \leq 1$, se sigue que $0 \leq -\hat{p}_{mk}\ log(\hat{p}_{mk})$. Se puede mostrar que el valor de la entropía será cercano a cero, si $\hat{p}_{mk}$'s son cercanos ya sea a cero o a uno, y al igual que el índice de Gini, la entropía tomará un valor pequeño en el m-ésimo nodo si éste es puro.

Al construir un árbol de clasificación, tanto el índice de Gini como la entropía son usados para calcular la calidad de una división en particular. Debido a que estos enfoques son más sensibles a la pureza del nodo que la tasa de error de clasificación. Cualquiera de estas tres aproximaciones puede ser usada cuando se haga la poda del árbol, pero la tasa de error de clasificación es preferible usarla si el objetivo es la precisión de la predicción del árbol podado final.

1.7.5.1. Árboles vs modelos lineales

Los modelos de regresión lineal asumen la forma

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{1.104}$$

mientras que un árbol de regresión asume la forma

$$f(X) = \sum_{m=1}^{M} c_m * 1_{(X \in R_m)} \tag{1.105}$$

donde $R_1, ..., R_m$ representan particiones del espacio de características.

Si la relación entre las características y la respuesta tienen una buena aproximación por medio de un modelo lineal como en (1.103), entonces el enfoque de regresión lineal podría funcionar bien y podría comportarse mejor que el modelo de árboles de decisión que no posé esta estructura lineal. Si existe una gran no linealidad y una compleja relación entre las características y la respuesta como indica el modelo (1.104), entonces la decisión de usar árboles de decisión podría mostrar un mejor comportamiento que el enfoque clásico.

1.7.5.2. Ventajas y desventas de los árboles

Ventajas

1. Los árboles son fácilmente explicables a las personas, inclusive quizá que explicar una regresión lineal, aunado al hecho de que puede ser representado gráficamente de manera sencilla.

2. Se puede pensar en árboles de decisión como una forma más cercana a como un humano tomaría la decisión, en lugar del enfoque de regresión o clasificación.

3. Los árboles pueden manejar predictores cualitativos de una manera más sencilla, sin la necesidad de crear variables auxiliares.

Desventajas

1. Los árboles generalmente no tienen el mismo nivel de precisión en la predicción comparado con otros modelos de regresión o clasificación.

2. Los árboles pueden no ser tan robustos, es decir, un pequeño cambio en la información puede causar un gran cambio en la estimación final del árbol.

Sin embargo, al agregar múltiples árboles de decisión, usar métodos como el *bagging*, *bosques aleatorios* y *boosting*, el comportamiento de la predicción puede ser mejorada notablemente.

1.7.6. Bagging

Los árboles de decisión vistos hasta ahora sufren de una gran varianza, lo que significa, que si se divide la información en dos partes de manera aleatoria, y ajustando un árbol de decisión para cada una de las mitades, se pueden obtener resultados completamente distintos. Por otro lado, un proceso con poca varianza, obtendrá resultados similares si se aplica a diferentes conjuntos de la información: por ejemplo, la regresión lineal tiende a tener poca varianza, si el ratio de n y p son grandes. El método de *Agregación Boodstrap* o *Bagging*, es un procedimiento cuyo propósito general es el de reducir la varianza del método de aprendizaje estadístico, por ello es frecuentemente usado en el contexto de árboles de decisión.

Una forma natural de reducir la varianza y por ende, incrementar la precisión de predicción del método estadístico de aprendizaje es tomar múltiples conjuntos de entrenamiento de una población dada, construir un modelo de predicción separado usando cada conjunto de entrenamiento y posteriormente promediar los resultados de las predicciones.

Definición 1.7.7. *(Bagging)*

Se define al método de Bagging como

$$\hat{f}_{\text{bag}}(x) \;=\; \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{b}(x) \tag{1.106}$$

Se puede calcular $\hat{f}^1(x), ..., \hat{f}^B(x)$ usando B conjuntos de entrenamiento por separado y promediando los de tal manera de obtener un solo modelo de aprendizaje estadístico con poca varianza.

Para aplicar el método de bagging a árboles de regresión, primero se construyen B árboles de regresión usando B conjuntos de entrenamiento bootstrapt y promediando dichos resultados. Estos árboles han de ser grandes y sin podar. Por lo tanto cada árbol individual tendrá gran varianza, pero un menor sesgo. Al promediar estos B árboles, se reducirá la varianza.

La forma más sencilla de aplicar el método de bagging a un problema de clasificación donde Y es cualitativo es, para cauna observación de prueba dada, se guarda la clase predicha por cada uno de los B árboles, y se toma la mayoría votada, la predicción general es aquella cuya clase sea la más común a lo largo de las B predicciones.

1.7.6.1. Error de estimación fuera de bolsa

Para estimar el error de prueba de del modelo bagged sin la necesidad de usar validación cruzada o el conjunto de validación.

Recordar que la clave para el bagging es que los árboles son ajustados repetidamente a través de subconjuntos de observaciones bootstrap, se puede observar que en promedio, cada árbol empacado usa alrededor de dos terceras pates de las observaciones. Al otro tercio de las observaciones no usadas para ajustar un árbol empacado se les conoce como observaciones fuera de la bolsa o OOB por sus siglas en inglés (out-of bag). Esto lleva a una única predicción OOB para la i-ésima observación.

La predicción OOB puede ser obtenida de esta forma para cada una de las n observaciones para OOB MSE general (en el caso de problemas de regresión) o bien el error de clasificación (para los problemas de clasificación). El error OOB resultante es un estimador valido del error de prueba para el modelo bagged, debido a que la respuesta para cada observación es predicha usando solo árboles que no han sido ajustados usando estas observaciones.

El enfoque de OOB para estimar el error de prueba llega a ser conveniente cuando se ejecuta el bagging en conjuntos de datos grandes en donde el método de validación cruzada puede a ser computacionalmente cansado.

1.7.6.2. Medidas de importancia variable

Cuando se empaca un gran cantidad de árboles, no es posible representar el resultado del aprendizaje estadístico con un solo árbol, y tampoco es claro cuáles variables llegan a ser las más importantes en este procedimiento.

Se puede obtener un resumen general de la importancia de cada uno de los predictores al utilizar la SRC (en caso del bagging de regresión de árboles) o bien el índice de Gini (en caso de bagging para clasificación de árboles).

En caso de árboles de regresión empacados, se puede registrar la cantidad total que la SRC (1.98) se reduce debido a la división hecha en un predictor dado, promediando sobre todos los B árboles. Un valor grande indica un buen predictor. De igual manera, en el contexto de clasificación de árboles empacados, se puede sumar la cantidad total que el índice de Gini se reduce al hacer divisiones en predictores dados, y promediarlos con todos los B árboles.

1.7.7. Bosques Aleatorios

El método de *Bosques aleatorios* o *random forest*, provee una mejora en el desempeño de árboles empacados al retocar la correlación de los árboles. Al igual que en bagging, se construye un número de árboles de decisión a partir de las muestras de entrenamiento bootstrap. Empero, cuando se construyen estos árboles de decisión, en cada momento que exista una división en el árbol, una muestra aleatoria de m predictores es elegida como candidatos de división del conjunto completo de p predictores. Esta división permite el uso de un único predictor, de los m seleccionados. Una nueva muestra de m predictores es tomada en cada división, y típicamente se elige $m \approx \sqrt{p}$, es decir, es el número de predictores considerados en cada división es aproximadamente igual a la raíz cuadrada del número total de predictores.

En otras palabras, en la construcción del bosque aleatorio, cada división en el árbol, al algoritmo no se le permite considerar la mayoría de los predictores disponibles.

Posteriormente en la colección de árboles empacados, la mayoría de todos los árboles usarán este predictor fuerte en la primera división. En consecuencia, todos los árboles empacados serán similares entre sí, por lo tanto, los predictores de estos árboles empacados estarán altamente correlacionados. Sin embargo, cuando se promedian tantos elementos altamente correlacionados, la reducción de la varianza es muy pequeña, lo que no pasaría si los árboles empacados estuviera menos correlacionados. Por lo tanto, el método de bagging no reducirá de manera sustancial la varianza sobre un solo árbol.

El bosque aleatorio ayuda a eliminar este problema al forzar a que cada división considere un solo conjunto de predictores. Por lo tanto, en promedio $(p-m)/p$ de las divisiones no serán consideradas como un fuerte predictor, y por ende, otros predictores tendrán una mayor oportunidad de serlo. Se puede pensar en este proceso como "descorrelacionar" los árboles, haciendo que el promedio de los árboles resultantes sea menos variable y más confiable.

La principal diferencia entre el método de bagging y bosques aleatorios es la elección de predictores de un subconjunto de tamaño m.

Si se usa un valor pequeño para m en la construcción de bosques aleatorios normalmente ayudará cuando se tenga un gran número de predictores correlacionados.

1.7.8. Boosting

Otro enfoque para mejorar los predictores resultantes de los árboles de decisión es el *boosting*.

El método de boosting funciona de manera similar al método de bagging, sin embargo la diferencia radica en que los árboles se crearán secuencialmente, cada árbol se construye usando información de árboles previamente construidos. Boosting no incluye muestras boodstrap, en su lugar cada árbol se ajusta en una versión modificada del conjunto original de datos.

Como el bagging, el boosting incluye combinar un gran número de árboles de decisión $\hat{f}^1, ..., \hat{f}^B$.

Este método, en lugar de ajustar un sólo gran árbol de decisión, cuyos montos para ajustar los datos duros y potenciar el sobre ajuste, el enfoque de boosting por otro lado "aprende lentamente".

Se ajusta el árbol de decisiones en relación a los residuales del modelo, en lugar del resultado Y, como respuesta. Posteriormente, se añade este nuevo árbol de decisión en la función ajustada para así actualizar los residuales. Cada uno de estos árboles pueden ser pequeño, con solo algunos nodos terminales determinados por el parámetro d del algoritmo. Al ajustar árboles pequeños a los residuales, lentamente se va mejorando $\hat{f}$ en áreas donde no se ejecuta de manera adecuada.

La clasificación de árboles a través de boosting, es similar pero un poco más complejo. Los detalles son omitidos en este trabajo.

Boosting tiene tres parámetros de ajuste:

1. B número de árboles. Boosting puede sobre ajustarse si B es muy grande, aunque este sobreajuste tiende a ocurrir lentamente, en caso de que ocurra. Se usa validación cruzada para seleccionar B.

2. λ, un número positivo pequeño. Este parámetro controla la tasa a la cual el método boosting aprende. Un λ muy pequeño puede requerir el uso de un mayor número B para poder tener un buen desempeño.

3. d es el número de divisiones de cada árbol, lo que controla la complejidad del arreglo boost. d es la profundidad de interacción y controla el orden de interacción del modelo boost, si existen d divisiones, implica que existen a lo más d variables.

Se muestra el error de prueba como función del número total de árboles y la profunda interacción de d.

Este modelo pose un mejor comportamiento que el modelo de bosques.

Al usar árboles pequeños se puede obtener también mejor interpretación de lo que está sucediendo.

1.7.8.1. Boosting para árboles de regresión

1. Sea $\hat{f}(x) = 0$ y $r_i = y_i \quad \forall i$ en el conjunto de entrenamiento.
2. Para $b = 1, 2, ..., B$ repetir:
   - Ajustar el árbol $\hat{f}^b$ con d divisiones ($d + 1$ nodos terminales) al los datos de entrenamiento (X, r).
   - Actualizar $\hat{f}(x)$ añadiendo en una versión menor del nuevo árbol

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{1.107}$$

   - Actualizar los residuales

$$r_i \leftarrow r_i - \lambda \hat{f}(x_i) \tag{1.108}$$

3. Resultado del modelo boost

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{1.109}$$

1.8. Vecinos más cercanos

Las funciones de distancia permiten identificar qué puntos están más cercanos a un cierto objetivo. Este principio es el que se encuentra detrás del Clasificador de Vecindarios Cercanos (KNN por sus siglas en inglés). Dado un conjunto de ejemplos previamente etiquetados, se busca el ejemplo de entrenamiento el cual sea más similar a un punto p no etiquetado. Una vez encontrado, la clase a la que pertenecerá p será aquella del conjunto de datos o vecindario etiquetado más cercano a éste.

Existen tres grandes ventajas de KNN para clasificación:

1. Simplicidad. El método de KNN no involucra una gran cantidad de matemáticas más allá de la medición de una distancia.
2. Interpretabilidad. Al estudiar los vecindarios cercanos se puede explicar por qué el clasificador tomó cierta decisión sobre el punto q de manera de precisa.

3. No linealidad. Este método tiene límites de decisión lineales a pedazos, es decir, a partir de un cálculo sabemos que las funciones lineales por partes se acercan a curvas suaves una vez que las piezas se vuelven lo suficientemente pequeñas. De esta manera, el KNN permite realizar límites de decisión muy complicados.

Un ejemplo análogo a este método en la vida real puede ser el usado ya sea por médicos o abogados para resolver algún caso. En ambos casos se revisa que información se tiene, si ha habido casos similares a ese en el pasado y que se utilizaron para tratarlo, de esta manera el médico aplicará los medicamentos necesarios para curar al paciente. En el caso del abogado, tomará los casos precedentes como base, así como la decisión hecha por los juristas previos para defender el caso.

Para clasificar un punto dado q, el método de vecinos cercanos regresa la etiqueta de q', el punto etiquetado más cercano a q. Ésta llega a ser una hipótesis razonable, asumiendo una similaridad en las características del espacio similar al espacio etiquetado.

Una clasificación más robusta es a través del voto múltiple de vecinos cercanos. Supóngase que se tienen k puntos cercanos al punto q, donde k es algún valor entre 3 y 50, dependiendo del tamaño de n. La disposición de los puntos etiquetados junto con la elección de k divide el espacio de características en regiones, con todos los puntos en una región dada en particular asignados a la misma etiqueta.

Cuando se tienen dos o más clases etiquetadas y se incrementa el tamaño de k, se tiende a producir una mayor región con límites más suaves, representando decisiones más robustas. Por otro lado, entre más grande se haga k, la decisión que se tomará será más genérica. Por ejemplo, $k = n$ es otra forma de nombrar al clasificador mayoritario, donde se asigna a q la etiqueta más común independientemente de sus características.

Para problemas de clasificación binaria, se desea que k sea un número impar, de tal forma que la decisión no se vuelva un empate. En general, la diferencia entre el número de votos positivos y negativos puede ser interpretado como una medida de confianza en la decisión.

Existe potencial asimetría relacionada a los vecinos cercanos. Cada punto tiene un vecino cercano, sin embargo para datos atípicos o aislados, estos vecindarios cercanos pueden no ser particularmente cercanos. Estos puntos atípicos pueden tener, de hecho, un importante papel en la clasificación al definir al vecindario con un mayor tamaño del que tienen realmente y llevar a una mala clasificación.

1.8.1. Encontrar el vecindario más cercano

Hay que tener en cuenta que este método es considerado como no paramétrico, pues el principio básico es contar cuántos miembros de cada clase se encuentran en cada conjunto cercano y regresa una fracción empírica estimada, es decir

$$p(y = c \mid q, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_k(x,\mathcal{D})} \mathbb{I}(y_i = c) \tag{1.110}$$

donde $N_k(\mathrm{q}, \mathcal{D})$ son los K vecindarios más cercanos de q en $\mathcal{D}$ y $\mathbb{I}(e)$ es la función indicadora definida como

$$\mathbb{I}(e) = \left\{ \begin{array}{ll} 1 & si\ e\ es\ verdad \\ 0 & si\ e\ es\ falso \end{array} \right\} \tag{1.111}$$

La medida de distancia más comúnmente utilizada es la distancia Euclidiana (lo cual limita la aplicación de ésta técnica a valores reales), sin embargo pueden existir otras métricas que pueden ser ocupadas.

1.8.2. La maldición de la dimensionalidad

El clasificador de vecindarios cercanos puede ser sencillo de aplicar si se cuenta con suficiente información de entrenamiento etiquetada. De hecho se puede demostrar que el clasificador puede tener uno de los mejores comportamiento si se encuentra en un factor de 2 y si $N \to \infty$. Sin embargo, el mayor problema con este clasificador es que no trabaja tan bien dentro de espacios de gran dimensión.

Esta "maldición" se puede explicar al considerar lo siguiente, supóngase que se aplica el método de vecindarios cercanos en cierta información que se encuentra distribuida uniformemente en una unidad cúbica de dimensionalidad D.

Supóngase que se estima la densidad de la clase etiquetada al rededor del punto q al crear un hipercubo alrededor de q hasta que contenga una cierta fracción de f puntos de información.

La distancia esperada de la orilla de este cubo sería $e_D(f) = f^{1/D}$. Si $D = 10$, y se quisiera tomar como base el 10 % de esta información, se tendría que $e_{10}(0.1) = 0.8$, por lo que se necesitaría extender el cubo un 80 % a lo largo de cada dimensión alrededor de q. Inclusive, si sólo se ocupara el 1 % de la información, $e_{10}(0.01) = 0.63$.

Dado que el rango total de la información es 1 a lo largo de cada dimensión, se puede apreciar que este método deja de considerar "vecindarios cercanos" y esto lleva a un problema

de predicción pues aquellos vecindarios que se encuentren lejos no serán buenos predictores de un punto dado.

Capítulo 2

Medición del desempeño, comparación y validación de clasificadores

2.1. Introducción

Evaluar el desempeño de un clasificador es una parte fundamental del proceso de modelado. Esta evaluación es importante para entender la calidad del clasificador, así como poder definir los hiperparámetros de manera adecuada, para que eventualmente puedan ser ejecutados con los modelos y así obtener los "mejores" resultados.

El proceso de evaluación de un modelo puede realizarse desde diferentes ópticas, algunas de manera gráfica, otras de manera teórica; y en esta última pueden dividirse entre pruebas paramétricas y pruebas no paramétricas.

En este capítulo se revisarán tanto las pruebas no paramétricas, las pruebas paramétricas, así como la evaluación gráfica a través de la curva ROC, además de hacer una revisión superficial a ciertas pruebas de independencia.

2.2. Evaluación del desempeño de un clasificador

Aunque hay varios criterios para evaluar el desempeño predictivo de los clasificadores, también son importantes otros criterios como la complejidad computacional.

2.2.1. Error de generalización

Típicamente, las medidas de desempeño predictivo son los principales criterios para seleccionar un clasificador. Sin embargo, las medidas de desempeño predictivo, tales como la exactitud (accuracy) se consideran como objetivas y cuantificables. Además, estás se pueden usar para tener algún algoritmo de referencia (benchmark).

Sea $C(D)$ el clasificador entrenado sobre el conjunto de datos D. El error de generalización de $C(D)$ es la probabilidad de clasificar erróneamente un punto seleccionado de acuerdo a la distribución del espacio de observaciones etiquetadas. La exactitud de clasificación es uno menos el error de clasificación.

Si se define al error de entrenamiento como el porcentaje de observaciones en el conjunto de entrenamiento que se clasificaron correctamente, i.e.

$$\hat{\epsilon}(C(D), D) := \sum_{(x,y)\in D} L(y, \hat{y}_C(x)), \tag{2.1}$$

donde $L(y, \hat{y}_C(x))$ es la función de pérdida cero-uno, que se define como

$$(y, \hat{y}_C(x)) = \left\{ \begin{array}{ll} 0 & si\ y = \hat{y}_C(x) \\ 1 & en\ otro\ caso \end{array} \right\} \tag{2.2}$$

la exactitud de clasificación es el criterio de evaluación principal.

Aunque el error de generalización es un criterio natural, su valor es desconocido ya que no se conoce la distribución de las etiquetas en el espacio de observaciones.

Se puede tomar al error de entrenamiento como una estimación del error de generalización. Sin embargo, utilizar el error de entrenamiento proporcionará un estimado optimistamente sesgado, especialmente si el clasificador sobre-ajusta a los datos de entrenamiento.

2.3. Esquemas de validación (Validación cruzada)

Como se dijo anteriormente, es difícil calcular directamente el error de prueba de un método de aprendizaje estadístico ya que no es común que se tenga un conjunto de prueba diseñado para esto y aún con esto, la distribución real de los datos es desconocida.

En este sentido, la validación cruzada es una técnica estadística que se puede usar para estimar el error de prueba utilizando los datos de entrenamiento sin contar con un conjunto de datos de prueba explícito.

Hay varias versiones de esta técnica, en todas se separa un subconjunto de las observaciones originales del procedimiento de ajuste del modelo y se aplica el método de aprendizaje a las observaciones que se excluyeron con el objetivo de evaluar su desempeño.

Se analizará tres versiones de validación cruzada:

- Metodología del conjunto de validación
- Validación cruzada leave-one-out
- Validación cruzada de k plieges.

2.3.1. Metodología del conjunto de validación

La metodología del conjunto de validación también se conoce como validación out-of-sample ó validación cruzada simple y como su nombre lo indica, se trata del método más sencillo de validación cruzada.

Bajo esta metodología se separa aleatoriamiente a las observaciones disponibles en dos partes de tamaño comparable, una de ellas será el conjunto de entrenamiento y el otro será el conjunto de validación.

El modelo se ajusta con el conjunto de entrenamiento y posteriormente se usa al modelo ajustado para hacer predicciones para las respuestas de las observaciones en el conjunto de validación.

Finalmente, se obtiene el *error de clasificación* sobre el conjunto de validación y se usa a éste como estimado del error de prueba.

Aunque este método es sencillo de implementar, tiene dos inconvenientes principales:

- Depende mucho de cómo se forman los conjuntos de entrenamiento y de validación, i.e. depende de la selección aleatoria. Por lo tanto, la estimación de validación del error de prueba puede diferir considerablemente.
- Como sólo un subconjunto de las observaciones en la muestra original están en el conjunto de entrenamiento y se usa éste para ajustar el modelo, con menos observaciones para entrenar el método de aprendizaje estadístico, la tasa de error de validación tiende a sobre-estimar a la tasa de error para el modelo ajustado en el conjunto de datos completo.

2.3.2. Validación cruzada Leave-one-out

La validación cruzada *leave-one-out* (LOOCV) es un refinamiento de la metodología del conjunto de validación. En vez de separar aleatoriamente al conjunto de observaciones originales en dos únicas partes, el LOOCV toma repetidamente al conjunto de entrenamiento como aquel que tiene a todas las observaciones disponibles excepto una y al conjunto de validación como simplemente la observación excluída.

De forma más específica, para las observaciones $\{(x_i, y_i)\}_{i=1}^n$, el LOOCV empieza usando a (x_1, y_1) como conjunto de validación y ocupa a las observaciones $\{(x_i, y_i)\}_{i=2}^n$ como conjunto de entrenamiento para ajustar el modelo de aprendizaje. Posteriormente se genera una predicción $\hat{y}_{(1)}$ usando el valor x_1 y $error_1 = I(y_1 = \hat{y}_{(1)})$ se usa como estimado del error de prueba.

Después, se repite el mismo procedimieto para la segunda observación, i.e. se toma a (x_2, y_2) como conjunto de validación y ocupa a las observaciones $\{(x_i, y_i)\}_{i=1, i\neq 2}^n$ como conjunto de entrenamiento para ajustar el modelo de aprendizaje. Posteriormente se genera una predicción $\hat{y}_{(2)}$ usando el valor x_2 y $error_2 = I(y_2 = \hat{y}_{(2)})$ se usa como estimado del error de prueba.

Se repite este procedimiento para cada una de las observaciones, i.e. se toma a (x_j, y_j) como conjunto de validación y ocupa a las observaciones $\{(x_i, y_i)\}_{i=1, i\neq j}^n$ como conjunto de entrenamiento para ajustar el modelo de aprendizaje. Posteriormente se genera una predicción $\hat{y}_{(j)}$ usando el valor x_j y se calcula $error_j = I(y_j = \hat{y}_{(j)})$.

Definición 2.3.1. *(Estimador de LOOCV del error de prueba)*

El estimador para validación cruzada de *leave-one-out* del error de prueba se define como

$$CV_{(n)} := \frac{1}{n}\sum_{j=1}^{n} error_j = \frac{1}{n}\sum_{j=1}^{n} I(y_j = \hat{y}_{(j)}) \tag{2.3}$$

La validación LOOCV tiene al menos dos ventajas con respecto a la metodología de validación cruzada simple:

(i) El LOOCV tiene menos sesgo. En LOOCV se aplica de manera repetida el método de aprendizaje en los conjuntos de entrenamiento, cada uno de estos conjuntos tiene $n-1$ observaciones, i.e. casi tantas observaciones como el conjunto de datos original. Por lo tanto, el LOOCV tiende a tener una tasa de error de prueba que tiene menor sesgo en comparación con la metodología del conjunto de validación.

(ii) Mientras la metodología del conjunto de validación depende críticamente de las observaciones en el conjunto de entrenamiento, la versión LOOCV no. Se puede decir que la separación entrenamiento/validación no tiene aleatoriedad y por lo tanto el estimado LOOCV del error de entrenamiento no muestra variabilidad.

Evidentemente, el precio que se tiene que pagar por estas ventajas en LOOCV es que la implementación y costo de cómputo puede ser alto ya que se tiene que ajustar el modelo n veces.

2.3.3. Validación cruzada de k pliegues

La validación cruzada de k pliegues (k-fold CV) es una metodología intermedia entre la validación cruzada simple y la LOOCV. Bajo ésta, se separa al conjunto de informaciones en k "pliegues" ó secciones de tamaño comparable entre sí. Se utiliza a las observaciones en un pliegue como conjunto de prueba, se entrena el modelo en las observaciones restantes y se obtienen predicciones para el conjunto de prueba. Se repite este procedimiento para cada uno de los pliegues.

Definición 2.3.2. *(Error de prueba en k-pliegues)*

El estimador *k-fold* del error de prueba se define como

$$CV_{(k)} := \frac{1}{k}\sum_{j=1}^{k} error_j \tag{2.4}$$

Donde cada error de prueba $error_1, \ldots, error_k$, corresponde al error respectivo de cada uno de los k-pliegues.

Claramente se puede observar que la LOOCV es un caso especial de k-fold CV cuando $k = n$.

Una ventaja evidente de k-fold CV es que si $k < n$ entonces es computacionalmente menos demandante (comparado con la LOOCV) obtener un estimado del error de validación.

Quizá una de las principales ventajas que tiene este tipo de validación es la relacionada con el *trade-off* sesgo-varianza (que se presenta muchas veces en el estudio de aprendizaje estadístico). El k-fold CV, con $k < n$, tiene el potencial de proporicionar estimados del error de prueba más precisos.

El LOOCV utiliza un mayor número de observaciones en el conjunto de entrenamiento para entrenar al método de aprendizaje estadístico, lo cual hace que el modelo que se escoja vía LOOCV tenga un sesgo bajo. Análogamente, validación cruzada simple usa un menor número de observaciones para entrenar al modelo por lo tanto es probable que el modelo que se seleccione vía validación cruzada simple tenga potencialmente un sesgo alto (en comparación con el que se obtuvo vía LOOCV). Por supuesto, en términos del sesgo del modelo seleccionado, la k-fold CV está en medio de LOOCV y la validación cruzada simple.

Sin embargo, este tipo de ganancias en aprendizaje estadístico no es gratuito. Dicha re-

lacción de orden en el sesgo, se traduce en una relación inversa para la varianza (entendida como la poca habilidad de ajustarse bien a datos nuevos que no se han observado).

Hablando de manera informal, la LOOCV lleva a un modelo con varianza grande pues $CV_{(n)}$ es un promedio de n cantidades altamente correlacionadas ya que cada una corresponde a casi las mismas observaciones. El grado de correlación es menor en k-fold CV y mucho menor en validación cruzada simple.

2.4. Esquemas de validación cruzada múltiple

Adicional a los métodos vistos previamente de validación cruzada, también existen los métodos de validación cruzada múltiple, dónde como su nombre lo indica, se repetirán en múltiples ocasiones. Sin embargo, la forma de hacerlo puede variar entre estos métodos.

Algunos de los métodos de validación cruzada múltiples son los siguientes.

2.4.1. Validación cruzada repetida

En K-Validación cruzda, se divide el conjunto de datos aleatoriamente en K pliegues y un modelo estadístico se reajusta K veces con los casos de cada pliegue, obtenidos a su vez, del conjunto de entrenamiento. Posteriormente, se procede a analizar la variación en el rendimiento de la predicción que resulta de elegir una división diferente de los datos. Una observación en relación a este método, es la falta de discusión formal en la literatura.

2.4.2. Validación cruzada estratificada

En una K-validación cruzada estratificada, la variable de salida se estratifica primero y el conjunto de datos se divide de forma aleatoria en K pliegues, asegurandose de que cada pliegue contenga aproximadamente la misma proporción de diferentes estratos. Sin embargo, cuando se ejecute este método, se tiene que cuidar que la estratificación no rompa la heurística de validación cruzada.

Con un gran número de validaciones cruzadas repetidas, la estratificación se vuelve redundante al seleccionar un modelo, mientras que para la evaluación de modelos, es aconsejable utilizar este método. Una observación en relación a este método es que no se tiene un consenso claro en cuanto a la aplicación de validación cruzada estratificada o cualquier estrategia de división que tenga en cuenta los valores de la variable de salida.

2.4.3. Selección de parámetros con repetición de cuadrícula

Una de las aplicaciones para la validación cruzada es el ajuste de parámetros e hiperparámetros en problemas de clasificación y regresión. En algunos casos es importante encontrar el K vecino cercano o bien el costo C de una máquina de soporte vectorial. Este método sugiere elegir un conjunto inicial de parámetros de entrada y realizar una validación cruzada de búsqueda de cuadrícula para encontrar parámetros óptimos (en relación a la cuadrícula y criterios dados). Este método puede generar estimaciones de rendimiento para cada uno de los puntos en la cuadrícula.

Se utiliza la validación cruzada de búsqueda de cuadrícula repetida, donde se hace repetición de N_1 veces la validación cruzada y para cada punto de la cuadrícula se generan N_1 errores de validación cruzada. Posteriormente se procede a elegir los parámetros de ajuste cuyo error medio de validación cruzada sea el mínimo, y nos referimos a él como la opción de validación cruzada óptima para los parámetros de ajuste.

2.4.3.1. Algoritmo de ajuste de parámetros a través de cuadrícula repetida de validación cruzada

Dado un conjunto de datos D, que consta de N elementos $(Y, X_1, X_2, ..., X_P)$, donde Y es la variable a predecir y $X_1, ..., X_p$ son variables predictoras. Se tiene un método de construcción de modelos F, para predicción de categorías, con un vector de parámetros de sintonía α. Se crea una cuadrícula de V puntos $\alpha_1, ..., \alpha_V$ y se quiere encontrar el valor óptimo entre ellos. Además, se tiene una función de pérdida *P()* como medida de bondad de ajuste.

1. Este proceso se ha de repetir N_1 veces
   - *a*) Dividir el conjunto de datos D aleatoriamente en K pliegues.
   - *b*) Para i desde 1 a K
     - Sea L el conjunto de D sin el i-ésimo pliegue
     - Sea T el conjunto i-ésimo de datos D
     - Para cada v desde 1 hasta V
       - Construir un modelo $f^v = f(L; \alpha^v)$
       - Aplicar f^v en T y guardar su predicción
   - *c*) Para cada valor α, calcular la bondad de ajuste con la función de pérdida *P()*, para topdos los elementos en D
2. Para cada valor α calcular la media de los N_1 valores de pérdida.

3. Sea α' el valor de α para el cual el valor promedio de pérdida es mínimo. En caso de existir múltiples α, se ha de escoger a aquella que pertenezca al modelo más simple.

4. Seleccionar α' como la opción de óptima de ajuste de parámetros a través de validación cruzada y de selección del modelo $f' = f(D; \alpha')$ como el modelo óptimo elegido a través de validación cruzada.

2.4.4. Doble validación cruzada

La validación cruzada se puede utilizar tanto para la selección y evaluación del modelo, sin embargo, requiere enfoques distintos. Por otra parte, se ha generado una tendencia a usar el error de validación cruzada como forma de evaluación del rendimiento del modelo, aunque al hacer esto se genera un sesgo en la estimación del error cuando se selecciona un modelo a través de validación cruzada. De ahí que surgan métodos tales como la "validación cruzada anidada" para generar una estimación imparcial del verdadero error.

La validación cruzada anidada, es la evaluación de validación cruzada del desempeño de una muestra grande de un modelo M elegido por un protocolo específico de validación cruzada R. Se define a R como la estimación del rendimiento de muestra grande del modelo M. Adicionalmente, si se estratifica la muestra a la que se aplicaría la validación cruzada, se ayudaría a reducir el sesgo de la estimación de la tasa de error resultante.

2.4.4.1. Algoritmo de validación cruzada estratificada anidada repetida

1. El protocolo de validación cruzada R consiste en repetir la K_1 validación cruzada N_1 veces, con una cuadrícula de V puntos $\alpha_1, ..., \alpha_V$. Designado por el modelo elegido M a través de la aplicación del protocolo de validación cruzada R.

2. Repetir el siguiente proceso N_2 veces

   a) Estratificar la variable de respuesta Y

   b) Dividir el conjunto D aleatoriamente en K_2 pliegues, asegurandose de que cada pliegue contenga la misma proporción de la variable Y estratificada.

   c) Para i desde 1 hasta K_2

   - Sea L el conjunto de D sin el i-ésimo pliegue
   - Sea T el conjunto i-ésimo de datos D
   - Aplicar el protocolo de validación cruzada para seleccionar el modelo f', es decir, repetir la validación cruzada K_1, N_1 veces con la cuadrícula de puntos $V, \alpha_1, ..., \alpha_V$ para encontrar un modelo elegido de validación cruzada óptimo f' en el conjunto de datos L.

- Aplicar f' en T

d) Calcular $P()$ para todos los elementos de D. Donde se ha de referir a este proceso como el error de validación cruzada anidado.

3. El intervalo entre el mínimo y el máximo de los errores de validación cruzada anidados de N_2 es el intervalo estimado R del error de muestra grande del modelo M. La media de dichos errores será la estimación R de lso errores de validación cruzada anidados N_2, del modelo M.

Una observación en relación a este método de validación cruzada, es que no existe una investifación formal que sugiera que el número de pleigues del bucle de validación cruzada externo K_2 y el interno K_1 deben de ser iguales o diferentes. De la misma manera, el número de repeticiones de la validación cruzada anidada puede o no ser igual al número de repeticiones de la validación cruzada externa.

2.4.5. Selección de variables y ajustes de parámetros

La relación entre la selección de variables y la validación cruzada fue aborada en primer instancia por Allen DM y Stone M, de manera independiente cada uno. Empero, esta selección fue hecha antes y no dentro de la validación cruzada. Hastie T, en el capítulo 7.10 de su libro *Los elementos del aprendizaje estadístico, Springer*, sefinen la forma correcta de llevar a cabo la validación cruzada, y es la siguiente:

1. Dividir las muestras en K pliegues de validación cruzada al azar.
2. Para cada pliegue $k = 1, 2, ..., K$

   a) Encontrar un subconjunto de variables predictoras "buenas" tales que muestren una correlación fuerte (univariable) con las etiquetas de clase, utilizando todas las muestras excepto las del pliegue k.

   b) Usando solo este subconjunto de variables predictoras, se ha de construir un clasificador multivadirado, usando todas las muestras excepto aquellas que se encuentran en el pliegue k, es decir, k^c.

   c) Usando el clasificador para predecir las etiquetas de clase para la muestra en el pliegue k.

   Los errores de estimación , serán acumulados sobre todos los K pliegues, para producir una estimación del error de predicción de la validación cruzada.

Al seleccionar variables y ajustar parámetros, el objetivo es seleccionar el número óptimo de variables, así como sus valores óptimos. De ahí, que es posible analizar este problema,

como si fuera un problema de optimización.

La validación cruzada entonces, sería utilizada para seleccionar el número de variables n y para hacer el ajuste de parámetros α de una cuadrícula multidimensional (n, α), donde $n \in (1, 2, ..., P)$ y $\alpha \in (\alpha_1, ..., \alpha_K)$. Para realizar esto, solo es necesario aplicar un ciclo de validación cruzada, pues trata a cada cuadrícula multidimensional de manera independiente. A continuación se muestra el algoritmo que llevará a cabo ésta tarea.

2.4.5.1. Algoritmo de validación cruzada a través de cuadrícula repetida para la selección de variables y ajuste de parámetros

Los siguientes pasos se han de repetir N veces.

1. Dividir el conjunto de datos D en K pliegues aleatoriamente
2. Para i de 1 hasta K
   - a) Sea L el conjunto de D sin el i-ésimo pliegue
   - b) Sea T el conjunto i-ésimo de datos D
   - c) Para r desde 1 hasta R
     - Sea $L' = S(L; r)$ donde L' es el conjunto L con únicamente r variables seleccionadas.
     - Sea T' el conjunto T con únicamente r variables seleccionadas.
     - Para v desde 1 hasta V
       - Construir un modelo estadístico $f' = f(L'; \alpha^v)$
       - Aplicar f' en T' y guardar la predicción
3. Para cada punto en la cuadrícula (n, α) calcular el promedio de pérdida.
4. Definir el par (r', α') con la mínima pérdida promedio como el par óptimo de números de variables seleccionadas y valores de parámetros.
5. Sea $D' = S(D', r')$, donde D' es el conjunto D con únicamente r' variables seleccionadas.
6. Seleccionar el modelo estadístico $f' = f(D'; \alpha')$ como modelo óptimo.

2.4.6. Doble validación cruzada

Es método incluye una validación cruzada interna para el ajuste de parámetros para cada conjunto de variables seleccionadas.

2.4.6.1. Algoritmo de doble validación cruzada

Este algoritmo consiste en dos pasos de validación cruzada. El primero.

1. Dividir el conjunto de datos D en K pliegues aleatoriamente
2. Para i de 1 hasta K_1
   - *a*) Sea L el conjunto de D sin el i-ésimo pliegue
   - *b*) Sea T el conjunto i-ésimo de datos D
   - *c*) Para r desde 1 hasta R
     - Sea $L' = S(L;r)$ donde L' es el conjunto L con únicamente r variables predictoras seleccionadas.
     - Sea T' el conjunto T con únicamente r variables predictoras seleccionadas.
     - Dividir el conjunto L' aleatoriamente en K_2 pliegues
     - Para v desde 1 hasta K_2
       - Se define a LL' como el conjunto L' sin el pliegue j-ésimo
       - Se define a TL' como el conjunto j-ésimo de L'
       - Para j desde 1 hasta K_2
         1) Construir un modelo estadístico $f' = f(LL';\alpha^v)$
         2) Aplicar f' en TL' y guardar la predicción
     - Para cada valor α, se tendrá que realizar el cálculo de $P()$ para todos los elementos de L'.
     - Se define a α' como el valor α cuya función de pérdida sea mínima.
3. Para cada número de variables seleccionado, calcular $P()$ para todos los elementos de D.
4. Definir r' como el número de variables selecionado, cuya función $P()$ es mínima.
5. Seleccionar r' como la validación cruzada óptima del número de variables seleccionadas.

El segundo paso

1. Sea $D' = S(D;r')$, dónde D' es el conjunto D con solamente r' variables seleccionadas.
2. Dividir el conjunto de datos D' en K pliegues aleatoriamente
3. Para i de 1 hasta K

a) Sea L el conjunto de D sin el i-ésimo pliegue

b) Sea T el conjunto i-ésimo de datos D

c) Para i desde 1 hasta K

- Sea L' como el conjunto de datos D' sin el i-ésimo pliegue
- Sea T' el conjunto i-ésimo del conjunto D.
- Para v desde 1 hasta V
  - Construir un modelo estadístico $f' = f(L'; \alpha^v)$
  - Aplicar f' en T' y guardar la predicción

4. Para cada valor α calcular la función de pérdida $P()$ para todos los elementos de D'.

5. Sea α' el valor de α cuyo valor de pérdida es mínimo.

6. Seleccionar α' como la elección óptima de validación cruzada para el ajuste de parámetros y $f' = f(D'; \alpha')$ como el modelo óptima de validación cruzada elegido.

En la literatura se llegan a usar indistintamente los términos validación cruzada doble y validación cruzada anidada, aunque no necesariamente impliquen lo mismo.

2.5. Calidad de un clasificador

Una vez que se seleccionó el esquema de validación, se puede construir el clasificador sobre los datos de entrenamiento y clasificar a dichas observaciones basadas en este modelo. Se necesitan medidas que describan qué tan bien se hizo la clasificación, éstas medidas pueden ser tanto discretas como probabilísticas y para ello han de tomar en cuenta tanto las clases reales como las clases estimadas para así poder determinar la calidad de los clasificadores.

2.5.1. Clasificadores discretos

Considérese un conjunto finito $X = \{x_1, \ldots, x_n\}$ de observaciones y para cada observación se conoce su clase real $y(x)$ y la clasificación $y_C(x)$ que devuelve el clasificador C. Como C es un clasificador discreto $y_C(x)$ toma valores en el mismo conjunto de clases $\mathcal{C} = \{C_1, \ldots, C_K\}$.

Una herramienta que es muy útil cuando se evalúan clasificadores discretos es la matriz de confusión M. La matriz de confusión M es una matriz de $K \times K$ (i.e. del número total de clases), en la que la entrada M_{ij} es el número de observaciones con etiqueta de clase real C_i pero que fue clasificada como C_j.

Ejemplo 2.5.1.

Ejemplo de un clasificador discreto c para un problema de clasificación multiclase.

x	$r(x)$	$c(x)$
x_1	A	B
x_2	B	B
x_3	C	A
x_4	A	A
x_5	B	B
x_6	B	B
x_7	B	A
x_8	A	C
x_9	C	B
x_{10}	A	A

En la segunda columna se muestra la clase real $r(x)$ correspondiente a la variable x dada por la primer columna. En la tercer columna $c(x)$ se muestra el valor dado por c para x.

En el caso de clasificación binaria, se tiene la siguiente notación:

- VP: Son los verdaderos positivos, es decir el número de observaciones que fueron predecidas con etiqueta T que efectivamente se observa etiqueta T.
- VN: Son los verdaderos negativos, es decir el número de observaciones que fueron predecidas con etiqueta F que efectivamente se observa etiqueta F.
- FN: Son los falsos negativos, es decir el número de observaciones que fueron predecidas con etiqueta F pero que en realidad son de la etiqueta T.
- FP: Son los falsos positivos, es decir el número de observaciones que fueron predecidas con etiqueta V pero que en realidad son de la etiqueta F.

Basándose en la matriz de confusión, se puede definir varias métricas.

Una de las medidas más conocidas es la exactitud de clasificación, que se denota por acc. Ésta se define como el cociente de las observaciones bien clasificadas, i.e.

$$acc(C) := \frac{1}{n} \sum_{k=1}^{K} M_k k. \tag{2.5}$$

Esta es una medida general que da una idea del desempeño global del clasificador, lo que implica que a valores más grandes, el desempeño será mejor. Al menos de manera global.

Otra medida muy conocida de evaluación, que sólo está definida para clasificadores binarios es el recall (que también se conoce como sensitividad o tasa de verdaderos positivos), i.e.

$$recall(C) := \frac{VP}{VP + FN} \tag{2.6}$$

en el denominador de esta expresión se encuentran todas las observaciones que tienen etiqueta V. Idealmente los valores de esta métrica debería de maximizarse para un mejor predicción.

Análogamente, se define a la precisión, denotada como $prec$, i.e.

$$prec(C) := \frac{VP}{VP + FP} \tag{2.7}$$

en el denominador de esta expresión se encuentran todas las observacion que se predijeron como V. Ai igual que en el recall, los valores de esta métrica deberían de maximizarse para una mejor predicción.

Otra métrica es la especificidad (que se conoce como tasa de verdaderos negativos), que se denota como $spec$, que es el número de observaciones correctamente clasificados como F, dividido entre el número de observaciones con etiqueta real F, i.e.

$$spec(C) := \frac{VN}{FP + VN} \tag{2.8}$$

en el denominador de esta expresión se encuentran todas las observaciones que tienen etiqueta F. El resultado obtenido de esta métrica será mejor, entre más grande sea, por lo que habrá que maximizarlo.

Finalmente, se define a la falsa alarma (que también se conoce como tasa de falsos positivos), que se denota por $falarm$, que es en número de falsos positivos comparado con el número de observaciones con etiqueta real F, i.e.

$$falarm(C) = \frac{FP}{VN + FP} \tag{2.9}$$

en el denominador de esta expresión se encuentran todas las observaciones que se predijeron como F. A diferencia de las métricas anteriores, el valor arrojado por falarm entre menor sea, será mejor.

Adicionalmente, se define a la medida F o al score F_1, que es la media harmónica entre la precisión y el recall, i.e.

$$F_1(C) \;=\; \frac{2}{recall^{-1} + prec^{-1}} \;=\; 2\left(\frac{prec + recall}{prec \cdot recall}\right). \tag{2.10}$$

Dado que F_1 depende de dos métricas iniciales, se deberá de maximizar el valor de este cálculo para tener un resultado balanceado y obtener mejores resultados en la predicción de los modelos.

Observación 2.5.1.

Teniendo en cuenta que si $F_1(C)$ se puede expresar como

$$F_1(C) = \frac{VP}{VP + \frac{1}{2}(VP + FN)} \tag{2.11}$$

entonces, se puede generalizar este score F_1, mediante el score F_λ que se define como

$$F_\lambda := (1 + \lambda^2)\left(\frac{prec \cdot recall}{(\lambda^2 \cdot prec) + recall}\right) \tag{2.12}$$

es decir,

$$F_\lambda = \frac{(1 + \lambda^2)VP}{(1 + \lambda^2)VP + \lambda^2 FN + FP} \tag{2.13}$$

Mientras más grande el valor de λ se pone más énfasis en el recall. Mientras más bajo el valor de λ, más influencia tiene la precisión.

Es común utilizar los valores:

- $\lambda = 2$ que le da más ponderación al recall que a la precisión (poniendo más énfasis en los falsos negativos).
- $\lambda = 1/2$ que le da menos ponderación al recall que a la precisión (atenuando la inflluencia de los falsos negativos).

$\triangledown$

Todas estas métricas que se definieron para clasificación binaria, se pueden usar para problemas de clasificación con más clases.

Una práctica común es calcular la medida para cada una de las clases de manera separada, i.e. una v.s. el resto y posteriormente promediar las medidas para cada una de las clases.

Definición 2.5.1. *(Kappa de Cohen)*

Otra métrica que se usa en clasificación multiclase es la kappa de Cohen y se define como

$$\kappa(C) := \frac{n\sum_{k=1}^{K} M_{kk} - \sum_{k=1}^{K} M_{k*}M_{*K}}{n^2 - \sum_{k=1}^{K} M_{k*}M_{*k}}, \tag{2.14}$$

donde M_{*k} es la suma de las entradas en la columna k de M, M_{k*} es la suma de las entradas en el renglón k de M.

No hay una respuesta simple y directa para la pregunta de que métrica de evaluación usar. En general, no hay algún clasificador que sea óptimo para cada métrica de valuación.

Cuando se está en problemas de clasificación multi-clase, la accuracy es más que suficiente junto con un análisis de la kappa de Cohen. Cuando los problemas están poco balanceados, se debe tomar en cuenta a los scores F_λ para verificar si hay un buen balance entre el recall y la precisión.

Por ejemplo, cuando hay un alto costo relacionado con clasificar observaciones en la clase negativa, será problemático una alta falsa alarma.

2.5.2. Evaluación en clasificadores probabilísticos

Ahora se estudiarán las métricas para clasificadores probabilísticos, i.e. clasificadores en los que el resultado no es una clase sino una probabilidad de que la observación pertenezca a cada clase.

La métrica de evaluación más importante para clasificadores probabilistas es la que se relaciona con el análisis ROC (Receiver Operating Characteristics). Estas técnicas colocan a los clasificadores en el espacio ROC, que es un espacio bidimensional en el que en el eje horizontal se encuentra la tasa de falsos positivos y en el eje vertical la tasa de verdaderos positivos.

Un punto (x, y) en el espacio ROC representa un clasificador con tasa de falsos positivos x y tasa de verdaderos positivos y. Algunos puntos especiales son:

(i) El punto $(1, 0)$, que reprenta al peor de los clasificadores, aquel en el que se clasificó como V a todos con etiqueta real F y a ninguno de los que tenían etiqueta V como V.

(i) El punto $(0, 1)$, que reprenta al clasificador perfecto, aquel en el que se clasificó como V a ninguno con etiqueta real F y a todos los que tenían etiqueta V como V.

Un clasificador es mejor que otro si está situado más hacia en noroeste en el espacio ROC, i.e. tiene baja tasa de falsos positivos y alta tasa de verdaderos positivos.

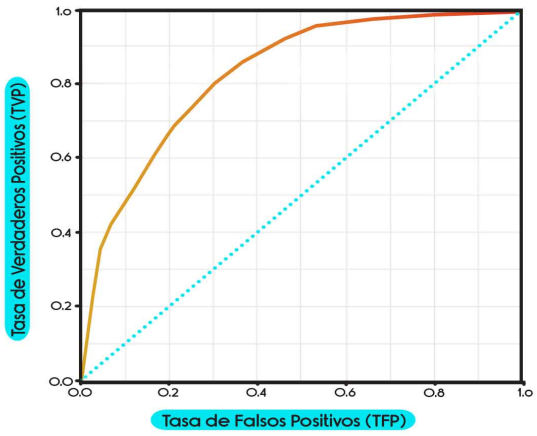


Figura 2.1: *Curva ROC*

Los clasificadores probabilísticos necesitan un umbral para hacer una decisión final para cada clase. Para cada uno de los posibles umbrales, se obtiene un clasificador discreto, con diferentes tasas de verdaderos positivos y de falsos positivos. Cuando se consideran todos los posibles umbrales y colocando a sus correspondientes clasificadores en el espacio ROC, se obtiene lo que se conoce como curva ROC.

Las curvas ROC son una gran herramienta de visualización para analizar el desempeño de los clasificadores. Una de las ventajas más importantes es que no dependen de la distribución de la clases, hecho que permite evaluar problemas no balanceados.

Con el fin de comparar dos clasificadores probabilísticos entre sí, se grafican las dos curvas ROC en el mismo espacio ROC.

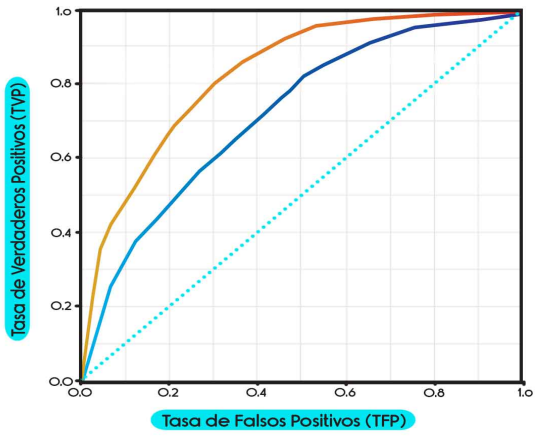


Figura 2.2: *Curva ROC. En este caso, la curva naranja se encuentra más cerca del punto (0,1), entre más cerca una curva se encuentre de este punto implica que su área es más cerca a* 1 *y tiene una mejor relación entre su especificidad y su sensibilidad. En este caso específico, la curva naranja representa a un modelo cuyo desempeño es mejor que el desempeño del modelo que representa la curva azul.*

Aunque es una herramienta de visualización muy útil, una métrica más objetiva es el AUC, Area-Under the Curve ó área bajo la curva.

El AUC es la superficie entre la curva ROC y el eje horizontal (del espacio ROC) y es una medida de la calidad del clasificador. Si lo ideal es que la curva ROC esté lo más al noroeste posible, se querá un AUC lo más grande posible, i.e. una superficie grande. Por supuesto, la interpretación de la curva ROC es más rica que simplemente el AUC.

Una precaución importante que se debe tener cuando se evalúa a los clasificadores usando curvas ROC, es que no se mide el desempeño absoluto del clasificador, si no el ordenamiento de las probabilidades. Por ejemplo en clasificación binaria, cuando las probabilidades de una clasificador son mayores que $\frac{1}{2}$ y con umbral también de $\frac{1}{2}$, ninguna observación se clasificará como negativa. Sin embargo, cuando todas las probabilidades de las observaciones positivas son mayores que las probabilidades de las observaciones negativas, el clasificador tendrá una curva ROC perfecta y el AUC será de 1. Es decir, la determinación del umbral para el clasificador final es importante.

Cuando se evalúan clasificadores probabilísticos multi-clase, una posibilidad es descomponer dichos problemas multiclase en varios problemas de dos clases y analizar la curva ROC para cada uno de los problemas binarios.

2.6. Comparación de clasificadores

2.6.1. Pruebas paramétricas

Primero se estudiará una comparación entre dos clasificadores y posteriormente la comparación entre varios clasificadores a la vez.

2.6.1.1. Comparaciones por pares

Un procedimiento clásico es la prueba t que se usa para comparar el desempeño de dos clasificadores (i.e. $k = 2$). La hipótesis nula es que el desempeño promedio de ambos clasificadores es el mismo y el objetivo es detectar una diferencia en las medias. Esta hipótesis nula se rechazará si la media de los desempeños es suficientemente diferente uno del otro.

Para ser más específico, se calcula la media muestral de las diferencias en el desempeño de los clasificadores, posteriormente se lleva a cabo una prueba t para evaluar si $\hat{D}$ es significativamente diferente de 0.

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n}(U_{i1} - U_{i2}) \tag{2.15}$$

La estadística de prueba es la media muestral studentizada

$$T := \frac{\bar{D}}{S/\sqrt{n}} \tag{2.16}$$

donde S es la desviación estándar de las diferencias.

Suponiendo que las diferencias tienen una distribución normal, las estadística T tiene una distribución t con $n - 1$ grados de libertad. La hipótesis nula se rechaza si $|T| > t_{n-1,1-\alpha/2}$, donde $t_{n-1,1-\alpha/2}$ es el cuantil al nivel $1 - \alpha/2$ de la distribución $t_{(n-1)}$.

2.6.1.2. Comparaciones múltiples

Cuando se comparan más de dos clasificadores se puede generalizar la prueba t a una prueba intra-sujetos de análisis de varianza (ANOVA).

La hipótesis nula es que el desempeño medio de los K clasificadores es el mismo y uno de los objetivos es detectar una posible desviación de esto. Dicha hipótesis nula se rechazará si los desempeños medios de al menos dos clasificadores son suficientemente diferentes entre sí. En el sentido de que la variabilidad media entre clasificadores (que se conoce como variabilidad sistemática) es suficientemente más grande que la variabilidad residual media (que se conoce como error de variabilidad).

Definición 2.6.1. *(Promedio total y promedio por región del clasificador)*

El promedio de un clasificador general, por columna o por renglón se define como:

$$\bar{Y}_{**} := \frac{1}{nK} \sum_{j=1}^{K} \sum_{i=1}^{n} Y_{ij}, \tag{2.17}$$

$$\bar{Y}_{i*} := \frac{1}{K} \sum_{j=1}^{K} Y_{ij}, \tag{2.18}$$

$$\bar{Y}_{*j} := \frac{1}{n} \sum_{i=1}^{n} Y_{ij} \tag{2.19}$$

Definición 2.6.2. *(Media de la variabilidad entre-clasificadores)*

La variabilidad entre-clasificadores media se define como

$$MS_{BC} = \frac{n}{K-1} \sum_{j=1}^{K} (\bar{Y}_{*j} - \bar{Y}_{**})^2 \tag{2.20}$$

y ésta mide las desviaciones medias cuadráticas del desempeño medio de los clasificadores con respecto al desempeño global medio.

Definición 2.6.3. *(Variabilidad media residual)*

La variabilidad media residual se define como

$$MS_{res} = \frac{1}{(n-1)(K-1)} \sum_{j=1}^{K} \sum_{i=1}^{n} [(Y_{ij} - \bar{Y}_{**}) - (\bar{Y}_{i*} - \bar{Y}_{**}) - (\bar{Y}_{*j} - \bar{Y}_{**})]^2 \tag{2.21}$$

Se puede decir que la variabilidad media residual es una medida de las variabilidad de lo que es realmente aleatorio.

La estadística de prueba está dada por

$$F = \frac{MS_{BC}}{MS_{res}} \tag{2.22}$$

Si esta cantidad es suficientemente grande, es muy probable que existan diferencias significativas entre los diferentes clasificadores.

Suponiendo que los desempeños siguen una distribución Gaussiana, la estadística F tiene una distribución $F_{(K-1,(n-1)(K-1))}$. Por lo tanto, se rechaza la hipótesis nula si $F > F_{(K-1,(n-1)(K-1)),1-\alpha}$, donde $F_{(K-1,(n-1)(K-1)),1-\alpha}$ es el cuantil $1-\alpha$ de la distribución $F_{(K-1,(n-1)(K-1))}$.

Si se rechaza la hipótesis nula, la única información que proporciona el ANOVA entrasujetos es que hay diferencias significativas entre el desempeño de los clasificadores pero no establece cuál o cuáles de los clasificadores tiene un desempeño sobresaliente.

2.6.2. Pruebas no-paramétricas

Una de las criticas principales de las pruebas paramétricas es que éstas dependen fuertemente de las suposiciones distribucionales subyacentes. En la práctica, no necesariamente se satisfacen, ó bien, no hay datos disponibles suficientes para verificar la validez de dichas suposiciones. En estos casos es recomendable usar pruebas no-paramétricas, que sólo requieren observaciones independientes.

En términos más formales, la hipótesis nula de las pruebas no-paramétricas que se estudiarán es que la distribución del desempeño de todos los clasificadores es la misma.

2.6.2.1. Comparaciones por pares

Cuando se compara dos clasificadores ($K = 2$), las pruebas más comunes son:

(i) Prueba de signos.

(ii) Prueba de Wilcoxon de signed-ranks.

Como bajo la hipótesis nula, los scores de los dos clasificadores son equivalentes, cada uno "ganará" en aproximadamente la mitad de los casos.

La prueba de signos rechazará la hipótesis nula si por ejemplo, la proporción de veces que gane el clasificador 1 es suficientemente diferente de $\frac{1}{2}$. Sea n_1 el número de veces que el clasificador 1 superó al clasificador 2. Bajo la hipótesis nula, n_1 sigue una distribución binomial con parámetros $(n, \frac{1}{2})$. Se rechazará la hipótesis nula si $n_1 < k_l$ ó $n_1 > k_u$, donde k_l es el entero mayor que satisface $\sum_{j=0}^{k_l-1} \binom{n}{j}(\frac{1}{2})^n \leq \alpha/2$ y k_u es el entero más pequeño que

satisface $\sum_{j=k_u+1}^{n} \binom{n}{j} (\frac{1}{2})^n \leq \alpha/2$.

Usando la aproximación binomial a la normal, si el número de observaciones es suficientemente grande, bajo la hipótesis nula, n_1 tiene aproximadamente una distribución normal con media $\frac{n}{2}$ y varianza $\frac{n}{4}$. Se rechaza la hipótesis nula si

$$\frac{|n_1 - \frac{n}{2}|}{\sqrt{n}/2} > z_{1-\alpha/2} \tag{2.23}$$

donde $z_{1-\alpha/2}$ es el cuantil al nivel $1-\alpha/2$ de la distribución normal estándar.

Una alternativa es la prueba de signed-ranks de Wilcoxon. Dicha prueba usa las diferencias $U_{i1} - U_{i2}$. Bajo la hipótesis nula, la distribución de estas diferencias es simétrica alrededor de la mediana y por lo tanto se debe tener que la distribución de las diferencias positivas es la misma que la distribución de las diferencias negativas. La prueba de Wilcoxon permite detectar alguna desviación de ésto para rechazar la hipótesis nula.

El procedimiento asigna un rango a cada diferencia de acuerdo al valor absoluto de estas diferencias, donde se asigna le media de los rangos en caso de que haya empates. Por ejemplo, si las diferencias son $0.03, 0.06, -0.03, 0.01, -0.04, 0.2$, los rangos respectivos son $3.5, 6, 3.5, 1, 5, 2$. Posteriormente, los rangos de las diferencias positivas y negativas se suman por separado. En este ejemplo, $R^+ = 12.5$ y $R^- = 8.5$.

Cuando hay pocas observaciones para rechazar la hipótesis nula, mín$\{R^+, R^-\}$ debe ser menor o igual que un valor crítico, dependiendo del nivel de significancia y el número de observaciones. Hay tablas para poder comparar estos valores. Por ejemplo, si $\alpha = 0.1$, el valor crítico es 2, lo que significa en este caso que la hipótesis nula no se rechaza a un nivel de significancia de 0.1.

Cuando sí hay observaciones suficientes, se puede utilizar una aproximación asintótica de la distribución de R^+ ó R^-. Sea $T = R^+$ ó $T = R^-$. Bajo la hipótesis nula, ambos tienen una distribución aproximadamente normal con media $\frac{n(n+1)}{4}$ y varianza $\frac{n(n+1)(2n+1)}{24}$. La prueba signed-rank de Wilcoxon rechaza la hipótesis nula si

$$\frac{|T - \frac{n(n+1)}{4}|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} > z_{1-\alpha/2} \tag{2.24}$$

donde $z_{1-\alpha/2}$ es el cuantil al nivel $1-\alpha/2$ de la distribución normal estándar.

2.6.2.2. Comparaciones múltiples

Las pruebas no-paramétricas que se estudiarán para hacer comparaciones múltiples de clasificadores son:

(i) Prueba de Friedman.

(ii) Prueba de Iman Davenport.

(iii) Prueba de rangos alineados de Friedman.

(iv) Prueba de Quade.

(v) Prueba de signos múltiples

2.6.2.3. Prueba de Friedman

Las pruebas de Friedman y de Iman Davenport son similares. Para cada observación, los clasificadores se ordenan de acuerdo a alguna medida de evaluación. Sea $r_i^{(j)}$ el correspondiente rango. Posteriormente se calcula el rango promedio R_j para cada clasificador a lo largo de todas las observaciones, i.e.

$$R_j = \sum i = 1^n r_i^{(j)} \tag{2.25}$$

El mejor método obtiene el rango 1, el segundo mejor obtiene el rango 2 y así sucesivamente.

Bajo la hipótesis nula, todos los clasificadores son equivalentes y por lo tanto todos los rangos de los diferentes clasificadores deben ser similares.

Tanto la prueba de Friedman como la de Iman Davenport buscan detencar una desviación de esto y pueden detectar si hay diferencias significativas entre al menos 2 de los métodos.

La estadística de prueba de Friedman está dada por

$$\tau = \frac{12n}{k(k+1)} \left(\sum_{j=1}^{k} (R_j - R)^2 \right) \tag{2.26}$$

Para un número suficiente de observaciones y clasificadores (como regla de dedo $n > 10$ y $k > 5$), y donde R_j es la media de los rangos de la columna j y R es la media general de los rangos o la media de todos los R_j, la versión alternativa a la estadística de Friedman y que normalmente simplifica los cálculas es:

$$\tau = \frac{12}{nk(k+1)} \left(\sum_{j=1}^{k} R_j^2 \right) - 3n(k+1) \tag{2.27}$$

τ tiene una distribución aproximada ji-cuadrada con $k-1$ grados de libertad. Se rechaza la hipótesis nula si $\tau > \chi^2_{k-1,1-\alpha}$.

En el caso de que se encuentre algún empate en los renglones, se les asignará el valor promedio de los rangos que están empatados. Debido a la forma en la que esta construida la prueba, es difícil que se encuentren múltiples empates, sin embargo en el caso de existir esta situación, se tendrá que tomar en cuenta un factor de corrección C, de tal manera que la estadística de Friedman quedaría establecida como:

$$\tau^* = \frac{\tau}{C} \tag{2.28}$$

siendo C igual a:

$$C = 1 - \frac{T}{n(k^3 - k} \tag{2.29}$$

Denotemos a t como el número de observaciones que contribuyen a un empate, de esta forma se tiene que $t^* = t^3 - t$, recordar que únicamente son empates por renglón y se repite el proceso de obtener los empates uno por uno. Una vez teniendo todos los t^* se puede obtener T que es la suma de todos los t^*.

2.6.2.4. Prueba de Iman Davenport

La estadística de prueba para la prueba de Iman Davenport, es

$$F = \frac{(n-1)\tau}{n(K-1) - \tau} \tag{2.30}$$

dicha estadística de prueba es más conservadora que la de Friedman y generalmente se prefiere. Si se tiene un número sufiecientemente grande de observaciones y clasificadores, F tiene una distribución aproximada $F_{(K-1,(n-1)(K-1))}$. Se rechaza la hipótesis nula si

$$F > F_{K-1,(n-1)(K-1),1-\alpha} \tag{2.31}$$

donde $F_{K-1,(n-1)(K-1),1-\alpha}$ es el cuantil al nivel $1-\alpha$ de la distribución $F_{K-1,(n-1)(K-1)}$.

Tanto en el caso de Friedman como en el de Iman Davenport, si la estadística de prueba es mayor que el correspondiente valor crítico, quiere decir que hay diferencias significativas entre los métodos de clasificación, pero no se puede hacer alguna otra conclusión (por

ejemplo cuál de los clasificadores es "mejor").

Ahora se estudiarán otras dos pruebas paramétricas que en ciertos escenarios pueden comportarse mejor que la prueba de Friedman, especialmente cuando el número de clasificadores es pequeño.

2.6.2.5. Prueba de Rangos alinados de Friedman

La prueba de rangos alineados de Friedman calcula los rangos de manera diferente. Para el conjunto de datos, se calcula el desempeño medio ó mediano de todos los clasificadores y se resta al desempeño de los diferentes clasificadores, a esto se le llama tener observaciones alineadas. Posteriormente, a las kn observaciones alineadas se les asigna un rango. Sea R_{ij} el rango del clasificador j para el dato i. La estadística de prueba de rangos alineados de Friedman es

$$T = \frac{(K-1)[\hat{R}_{*j}^2 - (Kn^2/4)(Kn+1)^2]}{[Kn(Kn+1)(2Kn+1)/6] - \sum_{i=1}^{n} \hat{R}_{i*}^2/K} \tag{2.32}$$

donde $\hat{R}_{*j} := \sum_{i=1}^{n} R_{ij}$ es igual al rango total del j-ésimo clasificador y $\hat{R}_{i*} := \sum_{j=1}^{K} R_{ij}$ es igual al ranfo total de la observación i.

Para un número suficientemente grande de observaciones, T tiene una distribución aproximada ji-cuadrada con $K-1$ grados de libertad. Se rechaza la hipótesis nula si $T > \chi^2_{K-1,1-\alpha}$.

2.6.2.6. Prueba de Quade

La prueba de Quade es una generalización de la prueba de rangos alineados de Friedman, incorporando el hecho de que no todas las observaciones son igualmente importantes. Es decir, algunas observaciones son más difíciles de clasificar que otras y los métodos que son capaces de hacer correctamente esta clasificación se deben favorecer.

La estadística de Quade calcula los rangos alineados basándose en el rando de los desempeños de diferentes clasificadores en cada observación. Es decir, primero se calculan los rangos de Friedman $r_i^{(j)}$ y posteriormente se calcula

$$\max_j U_{ij} - \min_j U_{ij} \tag{2.33}$$

y se ordenan de la misma manera que para la prueba de Friedman, en caso de existir empates deberá de ser el promedio de los rangos empatados. Sea Q_i el rango que se obtiene para la observación i. El rango ponderado ajustado promedio para la observación i con clasificador j es

$$S_{ij} = Q_i[r_i^{(j)} - \frac{K+1}{2}] \tag{2.34}$$

La estadística de prueba de Quade es

$$T_3 = \frac{(n-1)\sum_{j=1}^{K} S_j^2/n}{n(n+1)(2n+1)K(K+1)(K-1)/72 - \sum_{j=1}^{K} S_j^2/n} \tag{2.35}$$

donde $S_j := \sum_{i=1}^{n} S_{ij}$ es la suma de los rangos ponderados para cada clasificador.

La estadística T_3 tiene una distribución aproximada $F_{(K-1,(n-1)(K-1))}$. Se rechaza la hipótesis nula si $T_3 > F_{K-1,(n-1)(K-1),1-\alpha}$.

Cuando se rechaza la hipótesis nula, i.e. se rechaza el hecho de que todos los clasificadores tienen el mismo desempeño, los rangos promedio que se calculan para estos 4 métodos se pueden usar en sí mismos para ordenar qué método funciona mejor. Sin embargo, como en el caso de comparaciones múltiples de forma paramétrica, se siguen necesitandos procedimientos *post-hoc* para evaluar si las diferencias por pares son significativas.

La estadística de prueba para comparar el algoritmo j con el algoritmo l para las pruebas de Friedman e Iman Davenport es

$$Z_{jl} = \frac{R_j - R_l}{\sqrt{K(K+1)/6n}} \tag{2.36}$$

donde R_j, R_l son los rangos promedio que se calculan en los procedimientos de Friedman y Davenport.

Para el procedimiento de rangos alineados de Friedman, la estadística de prueba para comparar al algoritmo j con el algoritmo l es

$$Z_{jl} = \frac{\hat{R}_j - \hat{R}_l}{\sqrt{K(n+1)/6n}} \tag{2.37}$$

donde $\hat{R}_j = \hat{R_{*j}}/n$ y $\hat{R}_l = \hat{R_{*l}}/n$.

Finalmente, para el procedimiento de Quade, la estadística de prueba para comparar al algoritmo j con el algorimo l es

$$Z_{jl} = \frac{T_j - T_l}{\sqrt{\frac{K(K+1)(2n+1)(K-1)}{18n(n+1)}}} \tag{2.38}$$

donde $T_j = \frac{2}{n(n+1)} \sum_{i=1}^{n} Q_i r_i^{(j)}$ es el promedio ponderado de rangos que se describieron en el procedimiento de Quade.

Las 3 estadísticas Z_{jl} tiene una distribución aproximada normal estándar. Con este hecho, se puede calcular un p-valor, que es precisamente la probabilidad de que una variable con distribución normal estándar sea mayor que el valor absoluto de la estadística de prueba observada, i.e.

$$p - value = \mathbb{P}(Z > |z_{jl}|) = \Phi(-|z_{jl}|) \tag{2.39}$$

2.7. Pruebas de independencia

Es necesario antes de comenzar formalmente con el análisis, saber si los conjuntos de datos que se están utilizando son independientes unos de los otros. Para esto se implementarán dos pruebas que ayudarán a saber si la independencia realmente sucede entre los índices a ocupar. Estas pruebas serán la Ji-cuadrada y el Coeficiente de Correlación de Pearson.

2.7.1. Prueba de la Ji-cuadrada

Cuando la frecuencia del conjunto de datos consiste en categorías discretas, la prueba de la χ^2 puede usarse para determinar la importancia de la diferencia entre dos grupos independientes.

La hipótesis inicial de esta prueba suele ser que los dos grupos difieren respecto a algunas de sus características y por lo tanto, con respecto a la frecuencia relativa con la que sus miembros coinciden con las múltiples categorías. Para probar o rechazar esta hipótesis, se cuenta el número de casos por categoría donde los grupos tienen presencia; y se compara la proporción de casos por cateogría contra la proporción de casos del otro grupo en la misma categoría.

La hipótesis nula puede ser probada por:

$$\chi^2 \quad = \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{2.40}$$

donde:

- n_{ij} es el número de observaciones en los casos categorizados en el i-esimo renglón y en la j-ésima columna.
- e_{ij} es el número de casos esperados a clasificar en el bajo i-esimo renglón y en la j-ésima columna con H_0.

- $\sum_{i=1}^{r}\sum_{j=1}^{c}$ hace la suma directa sobre todas los r renglones y las c columnas, es decir, sobre todas y cada una de las celdas.

Los valores de χ^2 generados por la formula anterior, se distribuyen aproximadamente como una ji-cuadrada donde los grados de libertad l, son $l = (r-1)(c-1)$, donde r representa el número de renglones y c el número de columnas en la tabla de contingencia.

Para encontrar la frecuencia esperada para cada una de las celdas e_{ij}, se multiplican los dos totales marginales comunes a una celda en particular, y luego se divide este producto por el número total de casos N.

Si la frecuencia observada es muy cercana a la frecuencia esperada, la diferencia de $n_{ij} - e_{ij}$ serán por ende menores, y en consecuencia el valor de la χ^2 será menor y no se podrá rechazar la hipótesis nula. Por otro lado, si se tiene que las diferencias son grandes, entonces el valor de χ^2 será grande y cuanto mayor sea esta número, será más probable que los dos grupos difieran con respecto a las clasificaciones y pueda ser rechazada la hipótesis nula.

Cabe mencionar que existe una distribución muestra de diferencias para cada valor de *df*. Es decir, la importancia de cualquier valor particular de χ^2 depende del número de grados de libertad en los datos a partir de los cuales se calculo.

La prueba de la ji-cuadrada es aplicable a los datos en una tabla de contingencia solo si las frecuencias esperadas son lo suficientemente grandes, en caso contrario, si las frecuencias esperadas observadas no son tan grandes, se pueden aumentar los valores con la combinación de celdas, es decir, combinando clasificaciones adyacentes y reduciendo así el tamaño de celdas.

2.7.2. Prueba de Correlación de Pearson

Una medida de correlación es una variable aleatoria que expresa hasta que punto dos variables están relacionadas linealmente, normalmente estas dos variables pueden relacionarse en parejas de números.

Por definición, una medida de correlación entre dos variables X y Y deben de satisfacer las siguientes condiciones:

1. La medida de correlación debe tomar valores únicamente entre -1 y 1.

2. Si los valores más grandes de X tienden a estar apareados con los valores más grandes de Y, entonces los valores más pequeños de X y Y deben de estar apareados. A esto se le conoce como una medida de correlación positiva, cercana a 1.

3. Si los valores grandes de X tienden a estar apareados con los valores pequeños de Y y viceversa, entonces se dice que los valores X y Y tienen una correlación negativa, cercana a -1.

4. Si los valores de X se aparean aleatoriamente con los valores de Y, la medida de correlación deberá ser cercana a cero, esto debería de ocurrir en el caso de que X y Y sean independientes.

La medida de correlación de Pearson consiste en:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\ (Y_i - \bar{Y})^2\right)^{1/2}} \tag{2.41}$$

donde:

- El numerador es la covarianza muestral.
- El denominador es el producto de las desviaciones estándar.

El conjunto de datos a utilizar consistirá en una muestra bivariada de tamaño n, la cual se denominará como $(X,Y) = (X_1,Y_1),\ (X_2,Y_2),\ \ldots\ ,(X_n,Y_n)$. Se define a $R(X_i)$ como el rango de X_i, con $i = 1,...,n$, cuando se compara con el resto de los valores de X. De igual manera, se define a $R(Y_i)$ como el rango de Y_i, con $i = 1,...,n$, cuandos e compara con el resto de los valores de Y. En caso de existir empates, se asigna a cada empate el promedio de los rangos que serían asignados sino hubiese empates o repeticiones.

En caso de no existir empates, la correlación de Spearman (también conocida como ρ) se define como:

$$\rho = \frac{\sum_{i=1}^{n}\left(R(X_i) - \frac{n+1}{2}\right)\left(R(Y_i) - \frac{n+1}{2}\right)}{\frac{n(n^2-1)}{12}} \tag{2.42}$$

equivalente a

$$\rho = 1 - \frac{6\ \sum_{i=1}^{n}\left(R(X_i) - R(Y_i)\right)^2}{n(n^2-1)} \quad = \quad 1 - \frac{6T}{n(n^2-1)} \tag{2.43}$$

donde $T = \sum_{i=1}^{n}\left(R(X_i) - R(Y_i)\right)^2$.

Sin embargo, en caso de existir múltiples empates, es necesario ocupar la siguiente expresión:

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{i=1}^{n}(R(X_i))^2 - n\left(\frac{n+1}{2}\right)^2\right)^{1/2}\left(\sum_{i=1}^{n}(R(Y_i))^2 - n\left(\frac{n+1}{2}\right)^2\right)^{1/2}} \tag{2.44}$$

Dado que la prueba de ρ de Pearson es utilizara para probar independencia entre dos variables aleatorias de dos poblaciones, la prueba de hipótesis se puede dividir entres casos:

1. Caso 1, prueba de dos colas
   - H_0 Las X y Y son mutamente indepedientes.
   - Existe una tendencia para que los valores grandes de X estén emparejados con los valores más grandes de Y, de igual manera los valores chicos de X estén emparejados con los valores chicos de Y, o viceversa, los valores grandes de X con los valores pequeños de Y y los valores chicos de X, con los valores grandes de Y. En cualquier de los dos casos, $\rho \neq 0$.
2. Caso 2, prueba de una cola
   - H_0 Las X y Y son mutamente indepedientes.
   - Existe una correlación positiva entre los elementos de X y Y, es decir, $\rho > 0$
3. Caso 3, prueba de una cola
   - H_0 Las X y Y son mutamente indepedientes.
   - Existe una correlación negativa entre los elementos de X y Y, es decir, $\rho < 0$

De esta manera, usando a ρ como estadística de prueba, se tiene la siguiente regla de decisión:

- Caso 1, $\rho < \omega_{\alpha/2}$ ó $\rho > \omega_{1-\alpha/2}$
- Caso 2, $\rho > \omega_{1-\alpha}$
- Caso 3, $\rho < \omega_{\alpha}$

En el caso de que existiesen múltiples empates, será más conveniente usar la estadística T, en lugar de ρ. De esta manera, la prueba de ρ de Spearman, pasará a llamarse prueba de Hotelling-Pabst.

Dado que cuando T es grande, ρ es pequeño y vicebersa, la regla de decisión cambiará como se muestra a continuación, rechazando H_0 a un nivel de significancia de α, si:

- Caso 1, $\rho < \omega_{\alpha/2}$ ó $\rho > \omega_{1-\alpha/2}$
- Caso 2, $\rho < \omega_{1-\alpha}$
- Caso 3, $\rho > \omega_{\alpha}$

Capítulo 3

Construcción de índices e índices de precios en México

3.1. Introducción

De acuerdo con la Real Academia Española, un Índice puede significar una expresión numérica de la relación entre dos cantidades, es así, como un Índice bursátil es la expresión numérica que ayuda a los inversionistas a comparar el nivel de precios actuales de los activos financieros con los precios pasados, y de esta forma observar el comportamiento del mercado. Los índices bursátiles se calculan con base en diversos precios de activos seleccionados previamente.

Las dos principales características que tiene un Índice son:

1. Invertible, en el sentido que cualquier persona puede invertir en dicho índice bursátil al comprar un fondo de inversión indexado al éste.

2. Transparente, lo cual significa que cualquier persona puede revisar cómo está calculado el índice y reproducirlo.

A continuación se hará una revisión un poco más detallada de las formas en las cuales puede clasificarse un índice, teniendo en cuenta sus activos componentes, su región de origen, o bien la forma en la que éstos son calculados.

3.2. Tipos de índices por método de ponderación

La construcción de los índices puede variar por la propia construcción de la ponderación de los activos que los componen, es decir, en ciertos casos es posible encontrar índices que tengan en cuenta a los mismos activos bursátiles, sin embargo, la ponderación individual

de cada uno de éstos puede ser diferente. Algunos ejemplos de estos comportamientos son el S&P 500 y el S&P Equal Weighted, el segundo tienen una ponderación igualitaria para las 500 empresas que lo conforman. A continuación se mencionarán algunos de los métodos que son comúnmente utilizados para hacer la ponderación de activos dentro de un índice:

1. Capitalización de mercado. Este tipo de índices hace la ponderación de sus activos de pendiendo de la capitalización del mercado o el precio de sus acciones y el número de acciones en circulación, dividido por la capitalización de mercado total de todos los constituyentes del índice. Es decir,

$$w_i = \frac{(a_i)(v_i)}{\sum_{j=1}^{n}(a_j)(v_j)} \tag{3.1}$$

   donde, w_i es la ponderación del activo i, a_i es el precio del activo i y v_i es el volumen de operaciones del activo i. Y el denominador es la suma de los activos conformantes del índice por su volumen de operación.

   Este tipo de índices produce el mayor rendimiento para un nivel de riesgo determinado al ser evaluada bajo el modelo de *Capital Asset Pricing Model (CAPM)* o Modelo de valoración de activos financieros, el cual toma en cuenta la sensibilidad del activo o los activos al riesgo sistémico, el rendimiento esperado del mercado y el retorno de un activo libre de riesgo teórico.

2. Índices de libre flotación ajustada a ponderación por capitalización del mercado. Estos índices ajustan sus ponderaciones por las acciones en circulación que cada uno de los elementos conformantes mantienen estratégicamente y no están disponibles generalmente para el mercado público. Estas acciones pueden pertenecer a gobiernos, empresas del mismo grupo, afiliadas, fundadores de las empresas y en empleados. Los ajustes de flotación libre son empresas de gran complejidad y los diversos proveedores de índices tienen diferentes métodos de ajuste de flotación libre, por lo que a veces pueden producir resultados distintos.

3. Índices basados en ponderación por precio. Este tipo de índices ponderan sus activos por el precio de acción dividido por la suma de todos los precios de las acciones en el índice. Es decir,

$$w_i = \frac{(a_i)(p_i)}{\sum_{j=1}^{n}(a_j)(p_j)} \tag{3.2}$$

   donde, w_i es la ponderación del activo i, a_i es el precio del activo i y p_i es el precio del activo i. Y el denominador es la suma de los activos conformantes del índice por su precio en el mercado.

Para este tipo de índice se puede considerar como una cartera con un activo de cada uno de los activos constituyentes. Empero, en caso de que alguna entidad que posea activos dentro del índice divida sus activos, generaría que su ponderación en el índice disminuya; pues al dividirse los activos lo harán de igual manera el precio de éstos. Algunos de los índices que están conformados de esta manera son el Dow Jones Industrial Average y el Nikkei 225.

4. Ponderación equitativa. Para estos índices, todas las entidades que lo conforman poseen el mismo peso dentro del índice, es decir $1/n$ donde n es el número de integrantes del índice. Zeng Y Luo (2013) hace notar que estos índices son indiferentes a los factores y hacen aleatoria la fijación de precios errónea de los factores. Uno de los mayores problemas con este tipo de índices es que tienden a sobreponderar activos de poca capitalización; y a infraponderar los activos de mayor capitalización, en comparación con un índice ponderado por capitalización de mercado, llevando a los índices ponderados equitativamente a tener una mayor volatilidad y menor liquidez que aquellos índices ponderados por la capitalización del mercado. Un ejemplo de este tipo de índices es Barrons 400 Index, el cual está conformado por 400 entidades y donde a cada entidad se le asigna una ponderación igual a 0.25

5. Ponderación por factores fundamentales. Estos índices ponderan los activos que los constituyen con base de los factores fundamentales de las acciones, en lugar de los datos del mercado financiero de acciones. Es decir, estos factores fundamentales pueden incluir ventas, ingresos, dividendos y otros factores analizados en el análisis fundamental (de ahí su nombre) y que, al igual que este, supone que los precios del mercado convergerán a un precio intrínseco implícito en los atributos fundamentales.

6. Ponderación por factores. Estos índices ponderan a los elementos que los componen tomando en cuenta los factores de riesgo de mercado de cada uno de sus activos; medido en el contexto de modelos de factores tales como el *Fama-French three-factor model* o Modelo de tres factores de Fama y French. Algunos de los factores comunes que componen este grupo de factores pueden ser el crecimiento, el valor, el tamaño, rendimiento, calidad y volatilidad.

7. Ponderación por volatilidad. Constituyen índices cuya volatilidad relativa de precios es inversamente proporcional a su ponderación, es decir, a mayor volatilidad tendrá una menor participación dentro del índice. Dado que la volatilidad de los precios se define diferente por cada uno de los proveedores de índices, los métodos más comunes para el cálculo incluyen la desviación estándar de los últimos 252 días; y la desviación estándar semanal de los retornos de precios durante las últimas 156 semanas (alrededor de 3 años calendario).

8. Ponderación por mínima varianza. Son índices que usan un proceso de optimización de la varianza media. Esta clase de índices, castiga la ponderación que obtienen los

activos cuya volatilidad es alta y promueve a aquellos activos cuya volatilidad es mínima. Los activos altamente volátiles que estén correlacionados negativamente con el resto del índice pueden recibir ponderaciones relativamente mayores de las que se darían en el índice ponderado por volatilidad.

3.3. Tipos de índices por cobertura

Otro método de clasificación de los índices toma en cuenta al conjunto de activos de cobertura del índice. A este tipo de clasificación no le interesa cómo están ponderados sus elementos, si no las características económicas de dónde surgen sus elementos. Por ejemplo, el índice S&P Global 100 posé activos cuyo origen es global, por lo cual satisface a aquellos inversionistas que quieren tener sus inversiones diversificadas en todo el mundo. Otro ejemplo, aunque más regional, sería el MSCI Emerging Markets, el cual incluye acciones únicamente de países con un nivel de desarrollo económico similar (mercados emergentes), y satisfará a los inversionistas que deseen tener inversiones en países en vías de desarrollo, tomando en cuenta los riesgos que esto conlleva.

Las clasificaciones de este tipo de índices son los siguientes:

1. Cobertura por país. Representan el desempeño del mercado de valores de una nación determinada y por ende, es un gran indicador del estado de ánimo de los inversores ante la economía del país y las decisiones político-económicas que sus gobiernos tomen. Algunos ejemplos de este tipo de índices son el S&P 500 para Estados Unidos, el IPC para México y el DAX para Alemania.

2. Cobertura regional. Representan el desempeño del mercado de valores de una región geográfica determinada. En algunos casos pueden ser bloques económicos establecidos como en el caso del FTSE Developed Europe que representa a la Unión Europea, o bien el FTSE Developed Asia Pacific, el cual representa un conjunto de países localizados en Asia-Pacífico.

3. Cobertura global. Este tipo de índices, como su nombre lo indican, tienen acciones diversificadas a nivel global, por ejemplo, el S&P Global 100 y el FTSE Global Equity Index, conformado por más de 16,000 elementos.

4. Cobertura por intercambio. Como su nombre lo indican pueden basarse en el intercambio como el Nordic 40 o en grupo de intercambios como el Euronext 100.

5. Cobertura por sectores. Éstos rastrean el desempeño de sectores específicos de mercado. Pueden ser desde el sector tecnológico, el inmobiliario o el de la salud. Algunos ejemplos pueden ser el Nasdaq-100 enfocado principalmente a la tecnología o el índice de biotecnología Nasdaq.

3.4. Índice de Precios y Cotizaciones

El Índice de Precios y Cotizaciones (IPC) es el principal indicador bursátil de México, por lo tanto, el índice puede tomarse como un buen reflejo de la evolución del mercado accionario del país. Creado originalmente por la Bolsa Mexicana de Valores (BMV), busca medir el rendimiento de las 35 acciones más grandes y de mayor liquidez listadas dentro de la BMV.

Éste índice es calculado de manera diaria por Standard and Poor's (S&P) desde 2015, debido a una alianza estratégica entre ambas instituciones, esta alianza permite a S&P hacer una revisión del cálculo y la metodología de todos los índices generados por la BMV.

Teniendo en cuenta qué es un índice, cómo se conforman y cuál es su clasificación, se ahondará un poco más en la historia y la constitución de los índices de relevancia para este trabajo, los cuales son: el IPC, el INMEX y el S&P 500.

3.4.1. Un poco de historia sobre el IPC

El origen de este índice se remonta a 1900 cuando se creó el índice de "Promedio de Hechos" por parte de la Bolsa de Valores de México, antecesora a la Bolsa Mexicana de Valores (BMV), el cuál era calculado como la media aritmética anual de los valores operados por cada compañía listada. Sin embargo, por los problemas que acarreaba el frecuente cambio de las acciones listadas para 1958 se implementó un nuevo índice conocido como "Promedio de Cotizaciones de Acciones", el cual tomaba en cuenta sólo 11 acciones de empresas industriales y el cálculo del índice se basó en el precio promedio diario.

Conforme el mercado bursátil se fue expandiendo, la muestra de 11 acciones se fue volviendo obsoleta, por lo que para 1966 se decidió ampliar a 30 acciones, así como la modificación del cálculo para vincular el precio promedio con el valor anterior e introducir ajustes para eventos corporativos, tales como divisiones y fusiones (splits y contra splits).

Para 1978, la BMV crearía un nuevo índice que se estaría calculando a la par del Promedio de Cotizaciones de Acciones, este nuevo índice sería el IPC y permanecería como un índice privado, hasta 1980 cuando se sería revelado al público.

3.4.2. Cálculo del IPC

Dado que el IPC es un índice de capitalización de mercado, el cálculo de este se hará tomando en cuenta una "muestra" del global de acciones que cotizan en el mercado de la BMV; dicha muestra deberá de estar balanceada, ponderada y ser representativa del universo BMV. Esta muestra se selecciona semestralmente en los meses de marzo y de septiembre,

en un proceso conocido como "rebalanceo", el cual toma en cuenta el número de operaciones, importe negociado, días operados y un ratio entre el monto operado y el monto suscrito.

La ponderación de las 35 empresas que componen a este índice tiene como norma que ninguna de ellas deberá rebasar el 25 % de la ponderación dentro del índice, además, las principales 5 empresas en conjunto no podrán sobrepasar del 60 % de la composición total global del índice.

Dado que por mucho tiempo, fue el único instrumento que media al mercado accionario mexicano, existen instrumentos financieros que replican el comportamiento de este índice, tales como ETFs o TRACs.

3.5. Índice México

El Índice Mexicano (INMEX) es el segundo índice más representativo del mercado de valores mexicano, solo superado por el IPC. Al pertenecer ambos a México, comparten estructura, tamaño y necesidades del propio mercado bursátil mexicano.

Sin embargo, existen diferencias marcadas entre ambos índices, dentro de las cuales podemos encontrar:

1. Cantidad de acciones. A diferencia de las 35 acciones que conforman el IPC, el INMEX está conformado por un menor número de acciones, entre 20 y 25, las cuáles son las de mayor capitalización del sector.

2. Tamaño de los emisores. Para pertenecer al INMEX, las entidades emisoras de las acciones deberán de tener un valor de mercado mínimo de 100 millones de dólares. Requerimiento no necesario para pertenecer al IPC.

3. Ponderación de acciones. El peso de ponderación dentro del índice de cada una de las acciones integrantes no podrá superar el 10 % del valor del índice.

   Empero sus diferencias, también comparten algunas similitudes, entre ellas que el cálculo de este índice también se calcula de manera semestral y al ser ambos índices emitidos por la Bolsa Mexicana de Valores, también el Inmex es revisado por S&P Dow Jones .

   Dado que este índice toma en cuenta las compañías más grandes de la BMV, es posible hablar de sectores involucrados dentro de la composición del índice. Actualmente los sectores más grandes son consumo básico con un 33.2 %; sector de materiales con un 21.1 %; sector industrial con un 14.9 %.

3.6. Índice Standard & Poor's 500

Este índice, también conocido como S&P 500 o simplemente S&P, es un índice ponderado por capitalización que mide el desempeño de las acciones de las 500 empresas más grandes que cotizan en la bolsa de valores de los Estados Unidos; donde las 10 empresas más grandes representan el 26 % del total de capitalización del índice. Actualmente las empresas que conforman este índice son, Apple, Microsoft, Amazon, Facebook, Alphabet (Google), Berkshire Hathaway, Johnson & Johnson, JPMorgan y Visa, respectivamente. Este índice es recomendable si lo que se desea son inversiones con horizontes de tiempo a largo plazo.

Como curiosidad, al igual que el IPC y el INMEX, el S&P 500 es mantenido por S&P Dow Jones Índices, una empresa del grupo propiedad de S&P Global, y sus componentes son seleccionados por un comité experto cada cierto tiempo.

3.6.1. Un poco de historia del S&P 500

La historia de este índice se remonta a 1860, cuando Henry Varnum Poor estableció Poor's Publishing, empresa que publicó una guía para inversores de la industria ferroviaria. Posteriormente en 1923, Standard Statistics Company, fundada en 1906, comenzó a calificar bonos hipotecarios y desarrolló su primer índice bursátil compuesto por acciones de 233 empresas estadounidenses, donde su cálculo era semanal.

En 1941, Paul Talbot Babson compró Poor's Publishing y fusionó esta adquisición con Standard Statistics Company, convirtiéndose en la actual Standard & Poor's Corp.

Posteriormente, para 1957, el índice amplió el número de acciones a evaluar, pasando de 233 a 500 y el índice adquirió el nombre de S&P 500 Stock Composite Index.

Para 1976 The Vanguard Group, compañía administradora de inversiones con sede en Estados Unidos, ofreció el primer fondo mutuo a inversores minoristas que sigue el comportamiento del índice en el mercado.

Por último, en 2005 pasó a tener una ponderación por capitalización pública ajustada por flotación.

3.6.2. Cálculo del valor del índice

Los criterios para que el comité del S&P 500 seleccione o no una acción son: que tenga una capitalización de mercado de al menos 8.2 mdd, su liquidez, el domicilio donde se encuentra la empresa, la flotación pública, el estándar de clasificación global de la industria, la representación de las industrias en la economía de los Estados Unidos, la viabilidad fi-

nanciera y el periodo de tiempo que ha cotizado en la bolsa de valores.

El índice se recalibra cada tres meses, de manera que sea un indicativo fiel de las empresas más grandes de la Unión Americana, aunque el recalibrado no necesariamente implica que se haga una gran rotación de componentes del índice.

El S&P 500 es un índice ponderado por capitalización flotante, es decir, las empresas se ponderan en proporción a sus capitalizaciones de mercado, siendo su fórmula la siguiente:

$$\text{Nivel del índice} = \frac{\sum((p_i)(q_i))}{\text{Divisor}} \tag{3.3}$$

dónde p_i es el precio de la i-ésima acción en el índice, q_i es el número correspondiente de acciones disponibles públicamente ("flotante") para esta acción y el Divisor es un factor de normalización.

3.7. Comportamiento de los mercados financieros internacionales en los últimos años

Los mercados financieros se refieren en términos generales a cualquier mercado donde se realice la negociación de valores, dentro de los que se encuentran los mercados de, valores, bonos, divisas, derivados, etc. Todos estos elementos son de gran utilidad para el buen funcionamiento de una economía capitalista, de esta forma se pueden proveer de liquidez para empresas privadas, empresas públicas e inclusive los propios gobiernos.

Debido a su importancia dentro de las economías, es necesario tener en cuenta un panorama general de los mercados financieros, pues estos a su vez ayudará a entender el por qué los índices se comportaron de cierta manera en algunos periodos de tiempo. Hay que recordar que los índices se componen de acciones y estas acciones se comercializan dentro del mercado de valores, que a su vez, es uno de los componentes de los mercados financieros, por lo tanto, cualquier cosa que afecte a éstos, afectará en mayor o menor medida a los índices bursátiles.

A continuación, se muestra un resumen de los eventos más importantes que tuvieron lugar en los últimos años y que afectaron a los mercados financieros en cierta medida.

3.7.1. 2011

Durante el primer trimestre del 2011 se mantuvieron algunas tensiones derivadas de la recién generada crisis europea, ya que la situación fiscal en algunos países de la región era precaria, aunado a la incertidumbre que generaba la salud de algunos sistemas bancarios de la región. Debido a esto, algunos flujos de capital de economías emergentes mostraron

cierta volatilidad, sin embargo, para finales de marzo recobraron cierta estabilidad en dichos flujos, haciendo que se apreciaran las divisas de las economías emergentes.

Debido a lo anterior y al impacto menor que tuvieron los conflictos de Medio Oriente y el Norte de África, los mercados financieros internacionales generaron un entorno de menor aversión al riesgo.

Para el segundo trimestre del 2011, los mercados financieros internacionales se vieron afectados por un aumento en la incertidumbre, debido al aún grave problema fiscal y financiero de algunos países europeos y a que aún no se había llegado a un acuerdo con el congreso de los Estados Unidos para elevar el techo de endeudamiento, haciendo que se generara un temor a que se prolongara la debilidad de las principales economías avanzadas.

Para el tercer trimestre del año la situación no mejoró en términos de confianza, en razón a que los mercados internacionales mostraron una mayor volatilidad en sus precios. Esto fue consecuencia de la falta de acuerdos para resolver el tema de los desbalances en las economías avanzadas, así como la vulnerabilidad del sistema bancario de la zona Euro.

Por consiguiente, el ritmo de expansión de la actividad económica mundial siguió debilitándose, retroalimentándose negativamente por el efecto en los mercados financieros internacionales.

Llegado el cuarto trimestre del 2011, los mercados financieros tuvieron un comportamiento de gran volatilidad y aversión al riesgo inducido por la crisis del euro ya que persistieron los riesgos de deuda soberana así como la vulnerabilidad de gran parte de las instituciones financieras de esa región. Esta turbulencia financiera también afecto a los mercados financieros de las economías emergentes en los últimos meses del 2011, en particular los flujos de capital y el tipo de cambio, registrando una fuerte depreciación, aunado a una caída en los índices accionarios como el IPC.

Empero, a partir de las medidas adoptadas por el Banco Central Europeo (BCE), los mercados financieros mostraron una mejoría al aumentar los precios de los activos financieros en general, incluyendo los mercados accionarios y cambiarios de las economías emergentes.

3.7.2. 2012

Derivado de la turbulencia generada en el segundo semestre del año 2011, la economía mexicana tuvo una menor demanda externa y un aumento en las primas de riesgo de los mercados financieros, sin embargo, la mejoría en el entorno económico mundial que tuvo lugar en el primer trimestre del 2012 ayudó a reducir la probabilidad de algún evento ca-

tastrófico en los mercados financieros internacionales que a su vez, generó que se abatiera parte del referido incremento en las primas de riesgo.

Adicionalmente, las medidas implementadas por el BCE contribuyeron a disminuir la incertidumbre en los mercados financieros internacionales durante el primer trimestre y permitió que se reanudase el proceso de búsqueda, por parte de los inversionistas, de mejores rendimientos.

De esta manera, el flujo de capitales hacia las economías emergentes aumentó considerablemente en los primeros tres meses del 2012, en comparación con el mismo periodo de años anteriores.

Ya en el segundo trimestre del 2012, el ritmo del crecimiento de la economía global mostró un debilitamiento, principalmente debido a las economías de la zona Euro, así como las principales economías emergentes, a excepción de México.

En el ambiente de los mercados financieros, la incertidumbre persistió en gran medida debido a la volatilidad de los flujos de capital de las economías emergentes y esto se vio reflejado en sus mercados financieros respectivos.

Para el tercer trimestre del 2012, la economía mexicana comenzó a mostrar señales de desaceleración, sin embargo, al reducirse ligeramente la incertidumbre en los mercados financieros internacionales ayudó a mitigar un poco dicha situación.

Pese a las elecciones electorales mexicanas, el comportamiento de los mercados financieros tuvo un desempeño estable sin presentar grandes periodos de aversión al riesgo.

Durante el cuarto trimestre del año, la zona Euro logró acuerdos fiscales y monetarios que ayudaron a fortalecer la gobernanza de la región; por otro lado Estados Unidos hizo ajustes a sus políticas fiscales de corto plazo, lo que contribuyó a una mejoría en los mercados financieros internacionales, en particular, se observó un mayor apetito por activos de mayor riesgo, lo que se vio reflejado en los precios de los mismos.

3.7.3. 2013

En el primer trimestre de 2013, los mercados financieros internacionales se vieron favorecidos debido a las medidas adoptadas por las principales economías avanzadas para fortalecer la recuperación económica, en especial en la zona europea; sin embargo, se siguieron registrando ciertos periodos de volatilidad debido a las dificultades para enfrentar ciertos problemas fiscales y financieros de esa misma zona.

En el caso específico de México, al tener una economía integrada comercial y financieramente con el exterior, en especial con los Estados Unidos y debido a las medidas tomadas por este país, los mercados financieros nacionales se vieron beneficiados. Si bien los mercados financieros mostraron una cierta mejoría por las perspectivas de crecimiento de Estados Unidos, también persistieron los riesgos propios de la baja en el crecimiento de la economía mexicana.

Para el segundo trimestre de este 2013 debido a que en los Estados Unidos se fue recuperando lentamente el empleo, la inflación fue de acuerdo con lo esperado, se comenzaron a reducir los ritmos de compra de activos. Lo anterior generó un aumento en las primas de riesgo en este país, afectando a las economías globales y aumentando en la volatilidad de los mercados financieros internacionales.

Durante el tercer trimestre, los mercados financieros nacionales mostraron una elevada volatilidad, sin mayor repercusión al resto de la economía mexicana, lo anterior derivado de los ajustes económicos y fiscales realizados en este periodo.

Para finalizar el 2013, diversas economías emergentes enfrentaron un panorama económico complicado, ya que la acumulación de desbalances macroeconómicos exacerbó el impacto en la incertidumbre de los mercados financieros internacionales sobre los precios de los activos financieros.

En México, los flujos de capital continuaron mostrando una elevada volatilidad durante el último trimestre. Cabe señalar que ante el anuncio de la Reserva Federal de los Estados Unidos en diciembre de 2013, los mercados financieros nacionales se comportaron de manera ordenada.

3.7.4. 2014

Durante este primer trimestre del año, se tuvo un crecimiento moderado de la economía mundial, este crecimiento fue impulsado principalmente por el dinamismo de las economías avanzadas. En contraste, el ritmo de crecimiento de las economías emergentes disminuyó.

En Estados Unidos, la Reserva Federal continuó con el proceso de normalización de su política monetaria, llevando a los mercados financieros a mostrar un mejor desempeño para febrero, después de que en enero se presentara una gran incertidumbre.

En el segundo trimestre del 2014, el estimulado monetario realizado por las economías avanzadas ha generado una baja en la volatilidad de los mercados financieros y una recuperación de los flujos de capitales de las economías emergentes.

Para el tercer trimestre del año, la turbulencia en los mercados financieros aumentó de manera significativa teniendo periodos de gran volatilidad, lo anterior debido a múltiples motivos dentro de las cuales destacan la preocupación sobre el panorama económico mundial, la incertidumbre ante la respuesta de las economías avanzadas, así como los riesgos geopolíticos y la alarma sanitaria provocada por la pandemia del ébola. Los mercados de las economías emergentes en consecuencia fueron de los más afectados, sin embargo en el caso de México, los ajustes financieros se dieron de manera ordenada y bajo condiciones de adecuada liquidez.

En el último trimestre, los mercados financieros internacionales continuaron presentando periodos de gran volatilidad, acentuando las vulnerabilidades financieras de algunas economías emergentes. Lo anterior como consecuencia de la desaceleración de la economía mundial ante la debilidad que prevalece en la mayoría de las economías avanzadas y emergentes en general, exceptuando a los Estados Unidos.

3.7.5. 2015

Los mercados financieros internacionales durante el primer trimestre de 2015, mantuvieron una elevada volatilidad debido a la incertidumbre relacionada a la normalización de la política monetaria de Estados Unidos, además de la laxa política monetaria en otras economías avanzadas y la delicada situación económica en Grecia.

Aunque para el segundo trimestre del año, el aumento de la tasa de los fondos federales estadounidenses propiciaron el desarrollo de la economía mundial, en especial la de México, la incertidumbre sobre la normalización de la política monetaria de Estados Unidos aunado a: la precaria situación económica de Grecia, a los problemas en los mercados financieros chinos y la disminución de los precios de las materias primas, como el petróleo, contribuyeron a elevar la volatilidad en los mercados financieros internacionales. Esta alta volatilidad afectó los mercados nacionales, al observarse la exposición de riesgos del portafolio de inversionistas al demandar más coberturas cambiarias.

Para el tercer trimestre del año, continuó la incertidumbre alrededor del proceso de normalización de la política monetaria de los Estados Unidos. Adicionalmente, el crecimiento de la economía mundial continuó con niveles bajo y la economía china mostró indicios de desaceleración. Lo anterior contribuyó a que los mercados financieros internacionales continuaran con una alta volatilidad y una gran aversión al riesgo por parte de los inversionistas. Esto condujo a su vez a una caída en los precios de los activos financieros a nivel global, en especial las divisas de las economías emergentes y los índices accionarios en todo el mundo.

En el último trimestre de éste año, se realizó el primer ajuste al objetivo de la tasa de fondos federales de la Reserva Federal de los Estados Unidos, lo cual ayudó a disipar

momentáneamente la incertidumbre alrededor de los mercados financieros.

3.7.6. 2016

Durante los primeros tres meses del año, la actividad económica global siguió mostrando indicios de debilidad, hubo una reducción en el comercio mundial y se observó una nula recuperación en la actividad industrial estadounidense, desencadenando diversos episodios de volatilidad de los mercados financieros internacionales.

En el segundo trimestre del 2016, los mercados financieros internacionales mostraron un significativo aumento en la volatilidad y como resultado del referéndum del Reino Unido en torno a su salida de la Unión Europea. Sin embargo, debido a la pronta respuesta tanto del Banco de Inglaterra, como de otros bancos centrales de economías avanzadas, se lograron estabilizar temporalmente los mercados financieros internacionales.

Para el tercer trimestre del año y debido principalmente a los resultados de las elecciones de los Estados Unidos, los mercados financieros de todas las regiones registraron un incremento en la volatilidad, impactando especialmente al mercado mexicano, dada la gran relevancia de las políticas propuestas por el candidato Donald Trump sobre nuestro país.

Por lo que corresponde al último trimestre del año, prevaleció un ambiente de gran volatilidad como consecuencia, de la incertidumbre sobre el proceso de normalización de la postura monetaria de los Estados Unidos, así como el proceso electoral llevado a cabo en ese mismo país y el desenlace que se obtuvo. Lo anterior provocó un ajuste en los diversos portafolios de los mercados financieros internacionales e impactó de manera importante a los mercados nacionales, además de que dio lugar a una caída en los precios de los activos financieros y a un aumento significativo en la volatilidad.

3.7.7. 2017

El primer trimestre del 2017 y pese a la incertidumbre relacionada con la política económica y los crecientes riesgos geopolíticos, los mercados financieros internacionales mostraron una notable reducción en los niveles de volatilidad y el aumento en los precios de los activos respecto a lo observado durante el último trimestre del 2016. Respecto al desempeño de los mercados financieros nacionales, tuvieron afectaciones importantes al inicio del año principalmente como consecuencia de la incertidumbre en torno a las políticas comerciales y migratorias de la nueva administración estadounidense que podrían haber afectado de manera negativa a la economía mexicana.

Al igual que en el primer trimestre, en el segundo trimestre los mercados financieros mostraron una reducción importante en los niveles de volatilidad y en el aumento de los pre-

cios de los activos respecto al trimestre anterior. Esto se pudo deber a que durante estos dos periodos, se tuvieron expectativas de un escenario de crecimiento sostenido, apoyado por las condiciones crediticias favorables y la recuperación de las utilidades empresariales, así como la recuperación del comercio global, los mercados financieros mexicanos mejoraron durante este periodo pese a la incertidumbre de las políticas económicas y los riesgos geopolíticos.

Para el tercer trimestre, los mercados financieros internacionales se beneficiaron del escenario de recuperación económica y del incremento en la expectativa de que se aprobara un paquete fiscal en los Estados Unidos. Bajo este contexto, los precios de los activos financieros aumentaron tanto en economías avanzadas como en algunas economías emergentes. En particular, los índices accionarios registraron nuevos máximos históricos en algunas economías avanzadas.

Durante el último trimestre del año, la posibilidad del aumento en el ritmo de la normalización de la política monetaria estadounidense propició un entorno de mayor volatilidad en los mercados financieros internacionales.

3.7.8. 2018

En el primer trimestre del 2018, la posibilidad de sorpresas inflacionarias en algunas economías avanzadas, particularmente en los Estados Unidos, así como un escalamiento de las medidas proteccionistas y la materialización de algunos riesgos geopolíticos, generaron una mayor volatilidad en los mercados financieros internacionales y a su vez, un incremento en las tasas de interés y un menor apetito por el riesgo.

Durante el segundo trimestre del año, el comportamiento de los mercados financieros se mantuvo con una gran volatilidad así como un menor apetito por el riesgo, en consecuencia, los activos provenientes de economías emergentes presentaron un desempeño negativo. Así mismo, en México, se acrecentó la incertidumbre y la volatilidad asociada a la falta de acuerdos en el proceso de renegociación del Tratado de Libre Comercio de América del Norte y al proceso electoral que comenzaría a inicios del siguiente trimestre.

Para el tercer trimestre del año persisten algunos riesgos políticos y geopolíticos, además de que surgieron nuevos factores de incertidumbre, tales como las dificultades financieras presentadas por ciertas economías emergentes y el correspondiente riesgo de contagio. Lo anterior ha conducido a episodios de volatilidad en los diversos mercados financieros y en el comportamiento negativo de los activos financieros de las economías emergentes.

Para finales del 2018, el último trimestre presento un nuevo factor de incertidumbre en los mercados nacionales y fueron las políticas de la nueva administración estadounidense. Los mercados financieros nacionales mostraron periodos de gran volatilidad y aversión al

riesgo.

3.7.9. 2019

En el primer trimestre del 2019 los mercados financieros nacionales continuaron con periodos de alta volatilidad, así como aumentos en las primas de riesgo. Las perspectivas de pausa en el ritmo de normalización monetaria de las principales economías avanzadas propiciaron condiciones más favorables para las economías emergentes en los mercados financieros internacionales, lo cual en conjunto con las perspectivas de un marco macroeconómico sólido en México, contribuyó a un mejor desempeño de los mercados financieros nacionales.

Durante el segundo trimestre, se tuvieron perspectivas más moderadas en cuanto al crecimiento de la economía global, además persistieron algunos riesgos como las tensiones comerciales globales, principalmente relacionado con los Estados Unidos. Como fue el hecho a inicios de junio de ése año, la amenaza de imposición de aranceles a las importaciones de productos mexicanos por parte de los Estados Unidos, provocando un periodo de alta volatilidad para los mercados financieros nacionales.

Otro ejemplo de las tensiones comerciales provocadas por los Estados Unidos fue el anuncio a inicios de agosto, sobre la imposición de aranceles a las importaciones provenientes de China hacia ese país, generando episodios de volatilidad en los mercados financieros internacionales.

En el tercer trimestre del año, los mercados financieros internacionales presentaron relativa estabilidad al disminuir los niveles de volatilidad, resultado de la atenuación de las tensiones comerciales entre Estados Unidos y China y la reducción de la probabilidad de la salida desordenada del Reino Unidos de la Unión Europea.

En el ámbito nacional, debido al relajamiento de las condiciones financieras globales, se tuvo un mejor desempeño en los mercados financieros nacionales, lo anterior debido a que durante todo el año 2019 se presentaron prolongadas tensiones comerciales, mayores riesgos geopolíticos y factores idiosincrásicos, la economía global se fue desacelerando gradualmente.

Adicionalmente, para esta fecha comenzaron algunos brotes del nuevo coronavirus provenientes de China, generando una mayor volatilidad en los mercados financieros internacionales.

En relación a los mercados financieros de México, durante el cuarto trimestre del 2019, se registró un desempeño favorable. Empero, comenzaron a mostrarse ciertas señales de aversión al riesgo generadas por el nuevo brote de coronavirus.

3.7.10. 2020

Durante el primer trimestre del 2020, la epidemia del nuevo coronavirus se extendió a un gran número de países, incluyendo a México y se comenzaron a implementar medidas de confinamiento y distanciamiento social para contener la propagación del virus, debido a esto, se observaron múltiples afectaciones a la actividad económica mundial y a los mercados financieros.

Debido a la magnitud de las afectaciones, se comenzó a materializar una crisis económica global, la cual no tenía precedentes en las últimas décadas debido principalmente a su origen, tal como es el problema sanitario y no a un ciclo económico o financiero.

Lo anterior produjo una caída generalizada de los mercados financieros internacionales a mediados del mes de marzo, fecha en la cual se decretaron las medidas de restricción y confinamiento, siendo una de las mayores caídas en los índices accionarios de los que se tiene registro.

Aunado a lo anterior y derivado de las medidas implementadas, las cadenas globales de suministros se vieron interrumpidas, elevando por lo tanto la volatilidad de los mercados financieros internacionales.

Para el segundo trimestre del año, múltiples bancos centrales de economías emergentes decidieron recortar su tasa de interés para suministrar liquidez e impulsar el crédito y el buen funcionamiento de los mercados financieros, entre los cuáles se encuentra el Banco de México.

Los mercados financieros internacionales luego de sufrir fuertes pérdidas durante los meses de febrero y marzo, (momento en que se realizó el confinamiento más drástico), comenzaron a recuperarse gradualmente. Lo anterior como resultado de los estímulos monetarios, fiscales y financieros implementados por las economías de importancia sistémica, dieron como resultado un mayor optimismo por parte de los inversionistas ante la reapertura gradual de la actividad económica en diversos países publicando algunas cifras económicas donde se mostró una ligera recuperación para los meses de mayo y junio, con respecto al bimestre anterior.

A partir de la primer mitad de junio, los mercados financieros volvieron a mostrar ciertos periodos de volatilidad, resultado del temor de una segunda ola de contagios; y del incremento de casos de infectados en China, Estados Unidos y Japón, aunado a las discusiones sobre una posible sobrevaluación en los mercados accionarios a medida que se ha observado una mayor divergencia entre el comportamiento de los mercados financieros y de la actividad económica. Posteriormente, estos miedos fueron disminuyendo gradualmente y

por ende, la volatilidad disminuyó ligeramente.

Los principales índices accionarios, entre ellos el S&P 500, mostraron movimientos mixtos aunque acotados sin alcanzar todavía los niveles previos al inicio de la pandemia.

Durante el tercer trimestre del 2020, el comportamiento de los mercados financieros mejoró, aunque continuaron con periodos de volatilidad más acotada, impulsados por la gradual recuperación de la actividad económica mundial, los estímulos monetarios emitidos por los bancos centrales y los estímulos fiscales hechos por las principales economías avanzadas así como las expectativas por el desarrollo de una vacuna contra el covid-19.

Uno de los episodios que influyó en la volatilidad durante este trimestre, tuvo que ver con el proceso electoral de los Estados Unidos, así como la aprobación de un paquete fiscal en ese mismo país.

Ya para el último trimestre del 2020, los mercados financieros internacionales se vieron favorecidos principalmente por los avances en la producción y distribución de las vacunas contra el COVID-19, además de la finalización del proceso electoral de los Estados Unidos y las expectativas sobre las nuevas medidas contracíclicas a implementar por parte del nuevo gobierno estadounidense.

Por el contrario, el incremento en la volatilidad así como en las tasas de interés a largo plazo, tanto en economías avanzadas como emergentes y el aumento de la inflación, propiciado principalmente por los estímulos fiscales aplicados en las principales economías avanzadas, puede poner en peligro el crecimiento económico global y por ende la estabilidad de los mercados financieros internacionales.

Dentro del ámbito nacional, los mercados financieros se han visto gravemente afectados, así como la actividad económica, la inflación y las condiciones financieras generales del país, debido a las medidas que se han tenido que adoptar para la contención de la propagación de la pandemia.

Capítulo 4

Introducción al análisis técnico en finanzas

4.1. Introducción

Un análisis técnico es una herramienta o método, usado para predecir el futuro probable de un activo financiero, tomando como base la información del propio mercado.

La teoría detrás de la validez del análisis técnico es la noción de las acciones colectivas, compras y ventas, de todos los participantes del mercado que reflejan información relevante con respecto a un valor negociado, y posteriormente, se asigna continuamente un valor justo a dicho activo.

El análisis técnico es usado tanto por operadores técnicos del mercado tanto a nivel corporativo, como lo son fondos de inversión, fondos de pensiones, entre otros; y también por inversores individuales, a través de casas de bolsa.

Por otro lado, muchos operadores usan el análisis fundamental para determinar el clima donde sea propicio comprar en el mercado, pero habiendo hecho esa decisión, usan el análisis técnico para determinar con precisión, niveles de precios de entrada de compra con bajo riesgo.

4.2. Gráficos y ventanas de tiempo

Los operadores técnicos, aquellos que se basan en el análisis técnico, analizan los precios de las gráficas para predecir el momento de los precios. Las dos variables principales para el análisis técnico son las ventanas de tiempo consideradas y algunos indicadores técnicos, estos últimos pueden variar dependiendo de las preferencias del operador.

Las ventanas de tiempo utilizadas para el análisis técnico van desde el minuto a minuto

hasta ventanas mensuales o anuales. Las ventanas más populares elegidas por los operadores para analizar son:

- Gráfica de 5 minutos
- Gráfica de 15 minutos
- Gráfica de 60 minutos
- Gráfica de 4 horas
- Gráfica diaria

Las ventanas que cada uno de los operadores elige analizar dependerá del estilo personal de los propios operadores, por ejemplo, aquellos operadores que abren y cierran operaciones dentro del día de operación, prefieren analizar el movimiento de los precios en ventas pequeñas de tiempo, es decir, gráficas de 5 a 15 minutos; por otro lado, los operadores que mantienen posiciones de mercado de un mayor periodo de tiempo se inclinan más a analizar los mercados en ventanas más amplias, es decir, en gráficos que muestran 4 horas, ventanas diarias o inclusive semanales.

El movimiento de precios que ocurren dentro de periodos de 15 minutos puede ser significativo para operadores intra-día, los cuales buscan oportunidades para obtener una ganancia de las fluctuaciones de los precios que ocurren dentro del periodo de operación del mercado. Empero, ese mismo gráfico de movimientos de precios visto diariamente o semanalmente puede no ser significativo o no tener alguna importancia en los términos de operación a largo plazo.

Es fácil ilustrar esto al revisar el mismo precio del activo en gráficos de diferente temporalidad. La siguiente gráfica muestra la plata operada de manera diaria por varios meses, con un precio de \$16 dólares hasta los \$18.50. Un inversionista de largo plazo podría estar inclinado a buscar comprar plata basado en el hecho de que el precio se encuentra cercano a su punto más bajo.

Figura 4.1: *Gráfica diaria del precio de la plata, con gráfico de velas.*

Sin embargo, el mismo precio visto por en un gráfico de cada hora muestra una tendencia bajista que se ha acelerado en las últimas horas. Por lo tanto, un inversionista de plata interesado en realizar operaciones intra-día probablemente evitaría comprar el metal en ese momento.

Figura 4.2: *Gráfica intra-día del precio de la plata, con gráfico de velas.*

4.2.1. Gráfico de velas

Uno de los gráficos más utilizados dentro del análisis técnico para mostrar el precio de un activo es el gráfico de velas. Una "vela" se forma a partir de la acción del precio de un activo durante un periodo de tiempo único para cualquier periodo de tiempo, es decir, cada vela representada en una gráfica por hora mostrará la acción del precio del activo durante cada hora, por otro lado, en una gráfica de precios diaria, cada vela mostrará el comportamiento del precio del activo durante un día.

Figura 4.3: *Gráfica de precios diarios con representación de velas.*

Una vela se forma de la siguiente manera: el punto más alto de la vela, muestra el precio más alto al cual se operó un activo durante el periodo de tiempo establecido; de la misma forma, el punto más bajo de la vela indicará el precio más bajo al cual se operó ese mismo activo durante el tiempo establecido. El "cuerpo" de la vela (que normalmente se presenta en rectángulos color rojo y verde), indican los precios de apertura y cierre para el periodo de tiempo. Si el cuerpo de la vela es de color verde, esto indicará que el precio de cierre fue superior al precio de apertura, de lo contrario, el cuerpo de la vela será de color rojo.

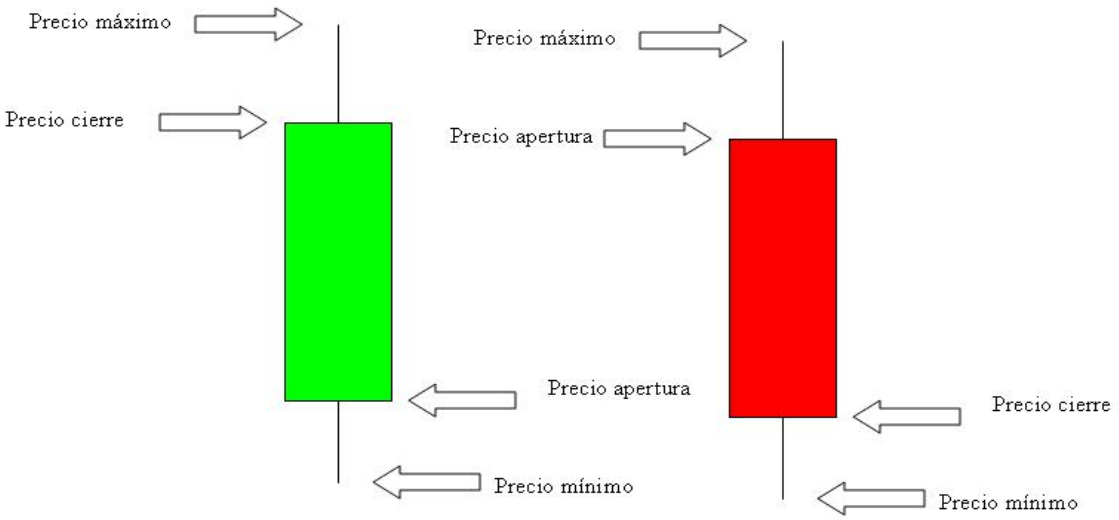


Figura 4.4: *Gráfica de velas explicada.*

4.2.1.1. Patrones de velas

Los patrones de velas que están formados por una sola vela o por una sucesión de dos o tres velas, son algunos de los indicadores técnicos más utilizados para identificar posibles reversiones del mercado o cambios de tendencia.

Por ejemplo, las velas Doji indican indecisión de un mercado que puede interpretarse como una señal de cambio de tendencia inminente o una reversión del mercado. La característica singular de una vela Doji es que los precios de apertura y cierre son los mismos, por lo que el cuerpo de la vela se representa como una línea plana. Cuanto más largas sean las "sombras" o "colas" superior o inferior para el periodo de tiempo, más fuerte será la indicación de indecisión del mercado.

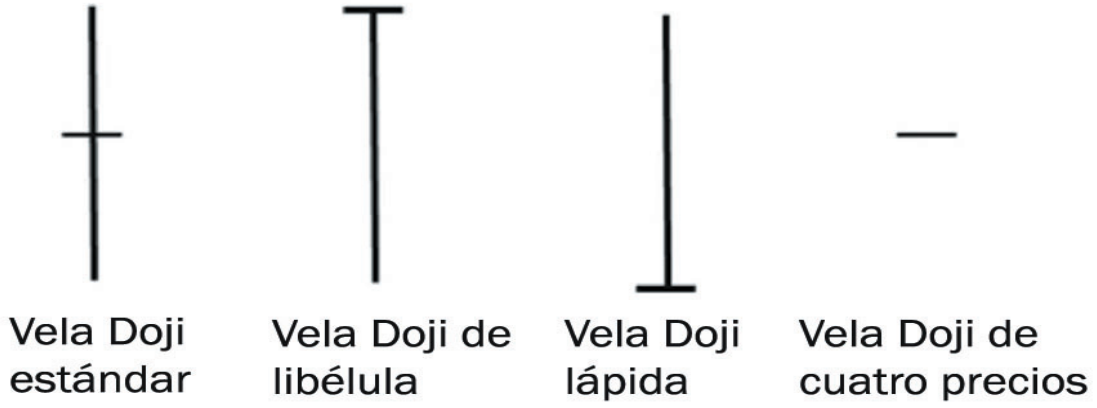


Figura 4.5: *Formas existentes de velas Doji.*

Una vela Doji típica es aquella de patas largas, donde el precio se extiende aproximadamente igual hacia cada dirección (superior e inferior). Cuando esta vela aparece después de una tendencia alcista o bajista extendida en el mercado, comúnmente se interpreta como una señal de un posible cambio de tendencia en dirección opuesta.

Por otro lado, los Doji de libélula aparecen después de una tendencia bajista prolongada, lo cual podría indicar un cambio de tendencia al alza. La libélula muestra a los vendedores presionando el precio aún más bajo (la cola larga es hacia abajo), pero al final del periodo, el precio se recupera para cerrar en su punto más alto. Esto se puede interpretar como que el mercado rechazo un empuje extendido a la baja.

Adicionalmente tenemos la gráfica Doji lápida (la cual tiene forma de cruz cristiana invertida), que hace referencia a una mala noticia para los compradores, esta vela es la opuesta a la vela de libélula pues indica un fuerte rechazo a un intento de hacer subir los precios del mercado, y por lo tanto, sugiere que podría seguir una posible tendencia a la baja.

Otra forma de Doji, es aquella de cuatro precios, e indica un movimiento de precios donde la apertura, el cierre y todas las transacciones del mercado sobre ese activo se realizan en el mismo precio exacto durante todo un periodo de tiempo. Este Doji indica una total indecisión por parte del mercado el cual no indica inclinación alguna hacia el alza o baja del precio del activo.

Aunque existen muchas otras formaciones de velas con diferentes nombres, las mencionadas son las más importantes y no se ahondará en el resto de las formaciones que existen.

4.3. Indicadores técnicos

Un indicador técnico es un cálculo matemático que puede ser aplicado a las series de tiempo generadas por los mercados financieros, por ejemplo, a la serie de precios, al volumen o inclusive a otro indicador técnico.

A diferencia del análisis fundamental, los indicadores técnicos no analizan el negocio perse, es decir, no se fija en los ingresos de una empresa, ni tampoco en su margen de ganancia.

Estos indicadores técnicos son usados por los operadores de mercado principalmente para analizar movimientos en el corto plazo dentro de los mercados primarios, para operaciones de largo plazo, estos indicadores llegan a proporcionar muy poca información.

Estos indicadores se pueden categorizar dependiendo del elemento del mercado que estudian, es decir, si analizan la tendencia, el impulso, la volatilidad o bien el volumen de operaciones sobre cierto activo.

4.3.1. Indicadores de tendencia

Como su nombre lo indica, estos indicadores miden la dirección y la fuerza de una tendencia comparando precios con una línea de base establecida, a continuación se hace mención a algunos ejemplos.

4.3.1.1. Medias móviles

Uno de estos recursos más utilizados son las medias móviles, los cuales actúan como suavizadores de los movimientos de los precios, ya sea a corto, mediano o largo plazo.

Se trata de una media del precio de un activo en un periodo de tiempo establecido, de este modo, la tendencia puede ser apreciada con más claridad, pese a que este indicador se forma con un cierto retraso y por ende no se anticipa al movimiento del mercado.

Existen diferentes tipos de media móvil:

- Simple. Una media móvil simple (MMS) es un promedio móvil aritmético calculado al añadir nuevos precios y dividirlos entre el periodo de tiempo abarcado en el cálculo del promedio. Los periodos cortos de tiempo, responden de manera más rápida a los cambios en el precio del subyacente, mientras que los promedios con un mayor horizonte de tiempo son más lentos en reaccionar.

  La MMS suaviza la volatilidad y hace notar más fácilmente la tendencia de los precios de un activo. Si la MMS repunta, significa que el precio del activo va en aumento, por el contrario si decrece, significa que el precio del activo va a la baja.

Su cálculo es el siguiente:

$$\text{MMS} = \frac{C_t + C_{t-1} + ... + C_{t_k}}{k+1} \tag{4.1}$$

Donde:

1. C_t es el precio del activo a tiempo t
2. k es el número de periodos a tomar en cuenta

- Ponderada. La media móvil ponderada, como su nombre lo indica, pondera con un mayor peso a aquellos datos que son más cercanos o relevantes al día de hoy, que aquellos que se encuentran más alejados en el tiempo. La suma de los pesos deberá de sumar 1 (o el 100 %). En el caso de la MMS, todos los pesos están igualmente distribuidos.

  El cálculo de este indicador es multiplicar el precio del activo por su peso asociado, sumar los resultados y dividirlo entre la suma de todos los pesos dados.

$$\text{MMP} = \frac{\sum_{t=1}^{n}(W_t)(C_t)}{\sum_{t=1}^{n} W_t} \tag{4.2}$$

  Donde:

  1. C_t es el precio del activo a tiempo t
  2. W_t es el ponderador del activo a tiempo t
  3. k es el periodo de tiempo a tomar en cuenta

  El primer elemento, deberá de tener una W_t mayor en comparación del último elemento, pues posee información más significativa en comparación.

Una estrategia cruzada podría ser comprar cuando el promedio móvil de periodo 10 cruza por encima del promedio móvil de 50 periodos.

4.3.1.2. Media móvil de Convergencia-Divergencia (MACD)

Es un indicador de tendencia que muestra la relación entre dos medias móviles del precio de un activo. El MACD se compone de calcular una media móvil exponencial (MME) de 26 unidades de tiempo, a partir de una MME de periodo 12.

Los inversionistas deberían de comprar el activo cuando el MACD cruce por encima de la línea señal y vender o estar corto cuando la línea sea cruzada en el extremo inferior.

El MACD tiene un valor positivo cuando el valor del MME de periodo 12 (MME(12)) se encuentra sobre el MME(26), y un valor negativo cuando el MME(12) este de bajo del MME(26).

A diferencia del Índice de Fuerza Relativa (RSI) que mide el cambio en el precio tomando en cuenta los máximos y mínimos, el MACD mide la relación entre dos médias móviles exponenciales (MME). Normalmente estos indicadores son usados en conjunto para proveer al analista de una imagen técnica más completa del mercado.

Su cálculo es el siguiente:

$$MACD(n)_{t-1} + \frac{2}{n+1}\left((MME(12) - MME(26)) - MACD(n)_{t-1}\right) \tag{4.3}$$

Cuanto mayor sea un número de promedio móvil, se considerará el movimiento de precios más significativo en relación a él. De esta manera, un promedio móvil de 50 o 100 periodos generalmente se considera mucho más significativo que el precio que se mueve por encima o por debajo de un promedio móvil de 10 periodos.

4.3.1.3. Pivotes, resistencias y soportes

Los indicadores de puntos de pivote diarios, que generalmente también ayudan a identificar varios niveles de soporte y resistencia además del punto pivote, son utilizados ampliamente por los operadores de mercado para identificar los niveles de precios para ingresar o cerrar operaciones. Si las operaciones se disparan, (o caen estrepitosamente), a través del pivote diario o todos los niveles de soporte o resistencia asociados, muchos operadores lo interpretaran como una operación de "ruptura" que cambiará los precios de mercado al alza o a la baja, dependiendo de la dirección de la ruptura.

En el caso del soporte, cuando el precio de un activo toca una zona de soporte, el operador deberá comprar dicho activo pues el precio tenderá a subir. Para que un nivel sea considerado como soporte tiene que haber una reacción de precios similar en el pasado donde éstos hayan rebotado al alza y también deberá acompañarse de un importante volumen de negociaciones de compra, ya que sin éste la fiabilidad del nivel de soporte no sería válida.

Por otro lado, cuando el precio de un activo toca una zona de resistencia, el operador deberá de vender dicho activo pues el precio tenderá a bajar. Para que un nivel sea considerado resistencia, se debe de haber actuado en el pasado como una zona de conflicto de precios al alza, siempre y cuando esté acompañado de un importante volumen de negociación de venta.

Una ruptura se define como el precio de un activo el cual rebasa un soporte o una resistencia, al alza o a la baja según sea el caso.

El cálculo de estos pivotes diarios y sus correspondientes niveles de soporte y ruptura ocupan los precios máximos, mínimos, de apertura y cierre del día anterior. La mayoría de los indicadores de puntos de pivote muestran el punto de pivote diario junto con tres soportes por debajo del punto pivote, así como tres niveles de resistencia de precio por encima de él.

4.3.1.4. Retroceso de Fibonacci

Leonardo de Pisa o también conocido como Fibonacci, fue un matemático italiano nacido en el siglo XII famoso por difundir en la Europa medieval el sistema de numeración con notación posicional, es decir, de base diez y el valor nulo cero.

Los índices Fibonacci, también llamados niveles, se utilizan comúnmente para identificar oportunidades comerciales y objetivos de ingreso comercial y ganancias que surgen durante tendencias sostenidas.

Las principales razones Fibonacci son $0.24, 0.38, 0.62$ y 0.76, los cuales se llegan a expresar en razón de porcentajes, es decir, $24\,\%$, $38\,\%$, etc. Cabe destacar que estos valores llegan a ser complementarios entre sí, i.e., el 0.24 es el complemento de 0.76; como el 0.38 es el complemento del 0.62; de tal manera que al sumarlos se forma un 1 o el $100\,\%$.

Al igual que con los pivotes, existe una gran variedad de indicadores técnicos disponibles para los operadores técnicos, entre ellos tenemos los "Retrocesos de Fibonacci", los cuales son utilizados para identificar puntos de entrada de comercio buenos y de bajo riesgo durante los retrocesos.

Un retroceso se define como un cambio de tendencia o rebote de activo después de que este haya mantenido una tendencia, alcista o bajista, por mucho tiempo; y se da antes de que el precio reanude la tendencia general a largo plazo.

Por ejemplo, suponga que el precio de una acción "R" ha subido de manera constante de $100 a los $200, posteriormente, el precio de las acciones comienzan a retroceder un poco. En este caso, muchos inversionistas buscarían un buen nivel de entrada para comprar acciones durante el retroceso de precios.

De esta manera, los niveles de Fibonacci sugieren que los posibles retrocesos de precios se mantendrán a una distancia igual de $24\,\%, 38\,\%, 62\,\%$ y $76\,\%$ del movimiento de tendencia alcista de $100 a $200. Los inversores observarán estos niveles en busca de indicios de que el mercado está encontrando soporte, nivel desde donde el precio comenzará a subir nuevamente. Es decir, los inversionistas esperarán a comprar acciones después de aproxi-

madamente un retroceso del 38 % en el precio de la acción, en este caso a un costo de $144.

4.3.1.5. Extensiones de Fibonacci

Continuando con el ejemplo anterior, si se compró la acción a $144 y se está tratando de determinar la ganancia objetivo al venderse todo, entonces se recurre a las extensiones de Fibonacci, método que indica que tan alto puede llegar a ser el precio de la acción cuando se recupere la tendencia alcista. Los niveles Fibonacci más comunes son: 1.26, 1.38, 1.62 y 1.76, los cuáles indican el precio futuro del activo. Ejemplo, hay que tener en cuenta que la diferencia entre el precio inicial y el final fue de $100 = $200 - $100, tomando esto como base se realiza el cálculo con el primer nivel de Fibonacci, 126 % = $144 + ($100 * 1.26) = $270, obteniendo como precio objetivo $270.

Aún cuando el operador no use la estrategia de Fibonacci o bien los pivotes de manera personal, valdrá la pena que los revise debido a que éstos son ampliamente usados por el mercado y podría haber una gran actividad comercial alrededor de estos puntos de precio, lo que podría ayudar al operador a obtener posibles movimientos futuros del mercado.

4.3.2. Indicadores de impulso

Este tipo de indicadores técnicos pueden identificar la velocidad del movimiento del precio comparando el precio de cierre actual con el de precio de cierre de sesiones anteriores.

4.3.2.1. Estocástico K % y D %

Un indicador estocástico es un indicador de impulso, desarrollado y aplicado a finales de los años 50 por George Lane, el cual compara un precio de cierre de un activo en particular con un rango de sus precios durante una cierta ventana de tiempo. La sensibilidad del oscilador a los movimientos del mercado se puede reducir ajustando la ventana de tiempo o tomando una media móvil del resultado. Se utiliza para generar señales comerciales de sobrecompra y sobreventa, utilizando un rango de valores entre $[0, 1]$.

Se manejan dos indicadores que son un oscilador estocástico rápido (%K) y un oscilador estocástico lento (%D), cuyas comparaciones son un buen indicador de la velocidad a la que los precios están cambiando o el impulso que tienen dichos precios.

La teoría general que sirve como base para este indicador menciona que en un mercado alcista, los precios cerrarán cerca del máximo, y en un mercado con tendencia bajista los precios tendrán un cierre cerca del mínimo. Las señales de transición se crean cuando el %K atraviesa una media móvil de tres periodos, a la cual se le denomina %D. De esta manera, el oscilador lento establece un periodo de desaceleración que controla el suavizado interno

de %K.

Debido a que se cree que el precio sigue el impulso, la intersección de la línea %K y %D se considera una señal de que puede estar en marcha una reversión, ya que indica un gran cambio en el impulso de un día a otro. Por otro lado, la divergencia entre el oscilador estocástico y la tendencia del activo también se considera una importante señal de reversión.

El cálculo del oscilador %K es el siguiente:

$$100\left(\frac{C_t - LL_{t-(n-1)}}{HH_{t-(n-1)} - LL_{t-(n-1)}}\right) \tag{4.4}$$

Donde:

1. $LL_{t-(n-1)}$ es el precio más bajo de los últimos $t-(n-1)$ días.
2. C_t es el precio de cierre del tiempo t
3. $HH_{t-(n-1)}$ es el precio más alto de los últimos $t-(n-1)$ días.

Para el cálculo del oscilador estocástico %D, se deberá obtener el promedio de los últimos tres días del oscilador estocástico %K.

Si se quiere utilizar una estrategia basada en el momentum, se deberá tomar posiciones largas en el activo el cual está teniendo tendencia positiva. Si por el contrario, la tendencia es bajista, entonces deberá de tomar una posición corta. Esta estrategia pretender vender barato pero comprar aún más barato o bien, comprar caro y vender aún más caro.

4.3.2.2. Larry Williams R

Es un tipo de indicador de impulso que se mueve entre 0 y 100 y mide los niveles de sobrecompra o sobreventa. Puede ser usado para encontrar puntos de entrada o de salida del mercado. El indicador es muy similar al oscilador estocástico y se usa de la misma forma. Fue desarrollado por Larry Williams y comprara los precios de cierre del mercado contra el rango que existe entre el mínimo y el máximo en el mismo periodo de tiempo. Comúnmente se ocupan 14 días para su cálculo, el cuál es el siguiente:

$$100\frac{H_t - C_t}{H_t - L_t} \tag{4.5}$$

Donde:

1. H_t es el precio máximo a tiempo t.
2. C_t es el precio de cierre al tiempo t.

3. L_t es el precio mínimo a tiempo t.

El indicador muestra cuando el precio es relativo al punto más alto de los últimos 14 días (o el periodo establecido).

Cuando el indicador se encuentra entre 0-20 significa que el precio es sobrecompra, o cerca del precio máximo. Cuando el indicador se encuentra entre 80-100 indica que el precio está en sobreventa, o lejos del puto máximo en el rango de tiempo.

La diferencia entre el indicador Larry Williams con el oscilador estocástico %K es que el primero representa el nivel de cierre de un mercado frente al máximo más alto del periodo retroactivo, por el contrario, el oscilador estocástico %K ilustra el cierre de un mercado en relación con el mínimo más bajo.

4.3.2.3. Índice Canal de Comodidad (CCI)

Es un oscilador basado en el impulso que se utiliza para ayudar a determinar cuándo un vehículo de inversión está alcanzando una condición de sobrecompra o sobreventa. También se utiliza para evaluar la dirección y fuerza de la tendencia de los precios. Esta información permite a los operadores determinas si quieren ingresar o salir de una operación, abstenerse de realizar una operación o agregar a una posición existente. De esta manera, el indicador se puede utilizar para proporcionar señales comerciales cuando actúa de cierta manera.

$$\frac{M_t - SM_t}{0.015 D_t} \tag{4.6}$$

Donde:

1. M_t es el promedio de la suma del precio máximo a tiempo t, precio mínimo a tiempo t y precio de cierre a tiempo t. Es decir, $(H_t + L_t + C_t)/3$.
2. $SM_t = \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$
3. $D_t = \frac{\sum_{i=1}^{n} |M_{t-i+1} - SM_t|}{n}$

El indicador se mueve desde un territorio negativo o cercano a cero hasta por encima de 100, eso puede indicar que el precio está comenzando una nueva tendencia alcista. Una vez que esto ocurre, los operadores pueden observar un retroceso en el precio seguido de un repunte tanto en el precio como en el CCI para señalar una oportunidad de compra.

De igual manera cuando el indicador pasa de lecturas positivas o cercanas a cero por debajo de -100, entonces puede estar comenzando una tendencia bajista. Ésta es una señal para salir de los largos o comenzar a buscar oportunidades de cortos.

4.3.2.4. Índice de Fuerza Relativa (RSI)

Es un indicador de impulso usado en el análisis técnico el cual mide el cambio en los precios recientes para evaluar las condiciones de sobreventa o sobrecompra en el precio de un activo. El RSI se muestra como un oscilador, (línea que se mueve entre ambos extremos de los datos), y puede tener una lectura entre el 0 y el 100. La ganancia o pérdida promedio en el cálculo es el porcentaje de ganancia o pérdida promedio durante un periodo retrospectivo. La fórmula usa un valor positivo para la pérdida promedio.

El RSI aumentará conforme existan incrementos en los cierres positivos, y caerán conforme el número de cierres negativos aumenten. La segunda parte del cálculo suaviza el resultado, de tal manera que el RSI se encuentre únicamente entre 0 y 100.

Su cálculo es el siguiente:

$$100 - \frac{100}{1 + \left(\sum_{i=0}^{n-1} \mathrm{UP}_{t-1}/n\right) / \left(\sum_{i=0}^{n-1} \mathrm{DW}_{t-1}/n\right)} \tag{4.7}$$

1. UP es el mayor precio al alza en la ventana de tiempo dada.
2. DW es el menor cambio de precio a la baja en la ventana de tiempo dada.

4.3.2.5. Indicador Momentum

El momentum es un indicador bursátil de tipo oscilador que mide la velocidad en el movimiento de los precios respecto a n periodos, donde n dependerá del estilo de cada uno de los operadores del mercado; y que generalmente eligen los números 10 y 12 como los periodos a observar.

El cálculo de este indicador es el siguiente

$$C_t - C_{t-n} \tag{4.8}$$

1. C_t es el precio de cierre del día de hoy
2. C_t es el precio de cierre de n días antes

Su interpretación se basa en torno a una línea neutral y su interpretación técnica se basa en los cortes que realiza la línea del indicador con respecto al eje del cero. De esta manera, el indicador emite señales de cuándo comprar o vender de acuerdo a lo siguiente: si la línea de momentum corta al eje del cero hacia arriba, esto indicará que es momento de comprar, en caso contrario, si corta al eje del cero hacia abajo, indicará que es momento de vender las posiciones que se tengan.

4.3.3. Indicadores de volumen

Este tipo de indicadores técnicos miden la fuerza de una tendencia en función del volumen de activos negociados en el mercado.

4.3.3.1. Oscilador de Acumulación y Distribución (A/D)

El oscilador A/D es un indicador acumulativo que usa el volumen y el precio para evaluar donde el activo está siendo acumulado o distribuido. La medida busca identificar las divergencias entre el precio de las acciones y el flujo del volumen. Esto proporciona una idea de la fuerza de una tendencia. Si el precio está subiendo pero el indicador está bajando, esto indica que el volumen de compra o de acumulación puede no ser suficiente para soportar el aumento del precio y podría producirse una bajada del precio.

Su cálculo es el siguiente:

$$\frac{H_t - C_{t-1}}{H_t - L_t} \tag{4.9}$$

1. H_t es el precio máximo a tiempo t.

2. C_{t-1} es el precio de cierre al tiempo $t-1$.

3. L_t es el precio mínimo a tiempo t.

El multiplicador en el cálculo proporciona un indicador de qué tan fuerte fue la compra o la venta durante un periodo en particular. Lo hace determinando si el precio cerró en la parte superior o inferior de su rango. Luego, esto se multiplica por el volumen, por lo tanto, cuando una acción cerca del máximo del rango del periodo y tiene un volumen alto, eso resultará en un gran salto A/D. Si el precio termina cerca del máximo del rango pero el volumen es bajo, el A/D no subirá tanto. Si el volumen es alto pero el precio termina más hacia la mitad del rango, el A/D tampoco subirá tanto.

Los operadores del mercado que hagan análisis técnico, tendrán que tomar en cuenta más de un elemento para conformar su estrategia, además de definir si se desea ser un operador a corto, mediano o largo plazo. Ningún indicador por si solo cuenta con suficiente eficiencia para lograr tomar una decisión de mercado adecuada, es por ello que los operadores deberán de realizar sus análisis tomando en cuenta múltiples indicadores, y que éstos a su vez, analicen múltiples perspectivas del mercado, desde la tendencia, el volumen, las gráficas, etc.

4.3.3.2. Indicador de Balance de Volumenes

El índice Balance de Volúmenes (On-Balance Volume OBV) es un indicador técnico de impulso comercial que utiliza el flujo de volumen para predecir cambios en el precio de las acciones, además de que muestra el sentimiento de la multitud en relación a un activo y que puede predecir un resultado alcista o bajista. Este indicador fue desarrollado en 1963 por Joseph Granville.

Granville describe las predicciones generadas por el OBV como "un resorte que se enrolla con fuerza". Él creía que cuando el volumen aumenta bruscamente sin un cambio significativo en el precio de la acción, entonces el precio eventualmente saltaría hacia arriba o bien caería.

Para realizar el cálculo de este indicador es necesario el volumen del activo en el mercado y el precio de cierre.

$$\mathrm{OBV} = \mathrm{OBV}_{\mathrm{prev}} + \left\{ \begin{array}{ll} \text{volumen}, & \text{si el cierre} > \text{cierre}_{t-1} \\ 0, & \text{si el cierre} = \text{cierre}_{t-1} \\ \text{volumen}, & \text{si el cierre} < \text{cierre}_{t-1} \end{array} \right\} \tag{4.10}$$

- OBV es el nivel actual en el balance
- OBV_{prev} es el volumen anterior en el balance
- volumen refiere al último volumen de operaciones

El volumen en balance proporcionará un total acumulado del volumen de negociación de un activo e indicará si el volumen entra o sale de un activo. El OBV es el total acumulado de volumen (sea positivo o negativo) y para hacer su cálculo existen las siguientes tres reglas:

1. Si el precio de cierre de hoy es más alto que el precio de cierre de ayer, entonces el OBV actual será igual al OBV del día ayer más el volumen de hoy
2. Si el precio de cierre de hoy es más bajo que el precio de cierre de ayer, entonces el OBV actual será igual al OBV del día de ayer menos el volumen de hoy
3. Si el precio de cierre de hoy es igual al precio de cierre de ayer, entonces el OBV de hoy permance igual al de ayer.

La teoría detrás del OBV se basa en la distinción entre los inversionistas institucionales e inversores minoristas. A medida que los fondos mutuos y los fondos de pensiones comienzan a adquirir un activo que venden los inversores minoristas, el volumen puede aumentar incluso si el precio se mantiene relativamente estable. Eventualmente, la ley de oferta y demanda elevará el precio de dicho activo. Llegado ese punto, los inversores más grandes

comenzarán a vender y los inversores pequeños comenzarán a comprar.

A diferencia de otros indicadores, el valor cuantitativo real del OBV no es relevante en sí, pues llega a ser acumulativo, mientras que el intervalo de tiempo permanece fijo por un punto de partida dedicado, lo que significa que el valor numérico real del OBV depende arbitrariamente de la fecha de inicio del análisis realizado. Por otro lado, los operadores y analistas se fijan en la naturaleza de los movimientos del OBV a lo largo del tiempo, pues la pendiente de la línea OBV tiene todo el peso del análisis.

Es así como lo operadores se llegan a fijar en los números de volumen en el OBV para rastrear grandes inversores institucionales, en especial revisan las divergencias entre volumen y precio como sinónimo de la relación entre los grandes inversores y los minoristas, con la esperanza de mostrar oportunidades de compra frente a las tendencias incorrectas.

Reservas del indicador

Debido a que el OBV es un indicador adelantado, puede producir señales falsas al no indicar lo que realmente sucedió en términos de las señales que produce. Es por ello que este indicador deberá complementarse con indicadores rezagados para poder realizar un análisis adecuado.

Adicionalmente, tiene la desventaja de que si en un día el mercado llega a tener movimientos bruscos de volumen en algún activo, ya sea al alza o a la baja, esto puede repercutir en el indicador durante varios periodos más adelante.

4.3.4. Indicadores de volatilidad

Estos tipos de indicadores técnicos miden, como su nombre lo indica, la volatilidad que existe en el precio de los activos negociados en el mercado.

4.3.4.1. Bandas de Bollinger

Las Bandas de Bollinger son una herramienta definida por un conjunto de líneas de tendencia que trazan dos desviaciones estándar (tanto positiva como negativa) sobre un promedio móvil simple del precio de un activo.

Esta técnica fue desarrollada por John Bollinger, diseñadas para descubrir oportunidades que brindan a los inversores una mayor probabilidad de identificar adecuadamente cuando un activo esta sobrevendido o sobrecomprado.

Para hacer el cálculo de esta herramienta es necesario seguir los siguientes pasos:

1. Calcular el promedio móvil simple (MMS) del activo a tratar, normalmente se usa una MMS de 20 días. Recordar que esto se hace sobre los precios de cierre.

2. Calcular la desviación estándar del precio del activo.

3. Sumar dos veces la desviación estándar a la MMS del precio del activo para obtener la banda superior. De igual forma restar dos veces la desviación estándar a la MMS del precio del activo para obtener la banda inferior.

Este indicador muestra que tan sobrevendido o sobrecomprado se encuentra un activo, es decir, si los precios de un activo se encuentran cerca de la banda superior, esto indicará que el mercado ha sobrecomprado dicho activo. Por el contrario, si el precio se encuentra cerca de la banda inferior, esto indicará que el activo ha sido sobrevendido por el mercado. Bollinger publicó 22 reglas que se han de seguir si se quiere utilizar este indicador para operar un activo, sin embargo, no se han de mencionar en este trabajo.

Reservas del indicador

Al igual que todos los demás indicadores, éste no deberá de ser usado de manera independiente. De hecho, su creador sugiere utilizarlo en conjunto con dos o tres indicadores no correlacionados que brinden señales de mercado más directas, tales como el indicador MACD o bien el RSI.

Otra limitante de este indicador es la forma de cálculo de la MMS, pues asigna el mismo peso a todos los datos, lo cual implica que puede perderse un poco de la información nueva con los datos más antiguos.

Adicionalmente, llega a ser arbitrario el como es elegido el periodo de 20 días para la MMS, así como el elegir dos desviaciones estándar hacia arriba o hacia abajo de la misma. Sin embargo, esto puede ajustarse al perfil de cada uno de los operadores del mercado, de acuerdo a sus respectivas estrategias.

Capítulo 5

Predicción de los movimientos de índices de precios en México

5.1. Introducción

Los datos que se utilizarán consisten de la dirección de movimiento de los precios de cierre diarios de los siguientes índices: Índice de Precios y Cotizaciones (IPC), del Índice México (INMEX) y del Standar & Poor's 500 (S&P 500).

El conjunto de datos abarca en un inicio del periodo del 25 de febrero del 2011 hasta el 02 de diciembre del 2020, i.e. 9 años y tres trimestres de información histórica. Sin embargo, una vez que se aplicaron los indicadores técnicos sobre los datos, resultaron algunas filas de información con "NA", principalmente el indicador de media móvil por lo que se eliminaron esos elementos resultando en menos datos a analizar, como se muestra a continuación:

(i) De 2448 observaciones iniciales para el IPC, resultaron 2422 finales.

(ii) De 2458 observaciones iniciales para el INMEX, resultaron 2430 finales.

(iii) De 2458 observaciones inicales para el S&P500, resultaron 2432 finales.

A continuación se muestran los históricos de dichos precios.

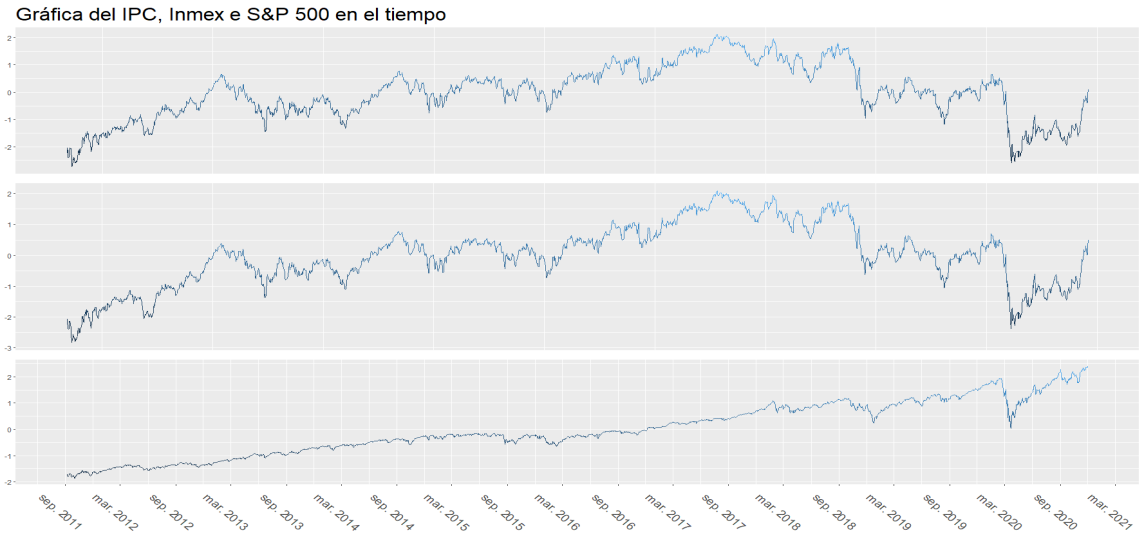


Figura 5.1: *Series de tiempo de los tres Índices a estudiar, la escala está normalizada.*

Para cada modelo se usarán 7 indicadores técnicos como variables explicativas (variables input ó predictores). Estos indicadores son ampliamente utilizados por administradores de fondos de inversión e inversionistas; y se consideran indicadores técnicos de señales de tendencia futura. Ya se habló de las ventajas de usar estos indicadores como variables explicativas en el Capítulo 4.

Los indicadores técnicos que se utilizarán son

(i) Promedio móvil simple a 10 días

(ii) Momentum

(iii) $K\,\%$ Estocástico

(iv) $D\,\%$ Estocástico

(v) Índice de fortaleza relativa (RSI)

(vi) Promedio móvil convergencia-divergencia (MACD)

(vii) Oscialador acumulador/distribución A/D

Las definiciones de dichos indicadores se estudiaron ampliamente en el Capítulo 3.

La dirección de los cambios que se dieron en el precio del índice se categorizarán como "subida" ó "bajada":

- Si el precio del índice al tiempo t es mayor que precio del índice al tiempo $t-1$ se considerará como "subida", i.e. si $S_t > S_{t-1}$.
- Si el precio del índice al tiempo t es menor que precio del índice al tiempo $t-1$ se considerará como "bajada", i.e. si $S_t < S_{t-1}$.

El conjunto de variables explicativas se estandarizará para que esté en una escala de $[-1, 1]$. Esto evitará que alguno de los atributos extremos (grandes y pequeños) en la escala original ejerza demasiada influencia numérica en la ejecución algorítmica.

5.2. Determinación de la arquitectura

Para cada una de las técnicas, se determinarán los hiperparámetros de cada una por un esquema de validación cruzada de 5 pliegues (5-cross-validation).

A continuación se enlistan las técnicas estudiadas en este trabajo así como sus hiperparámetros:

1. Regresión logística (sin hiperparámetros)
2. Regresión probit (sin hiperparámetros)
3. Regresión logística con regularización Ridge. Con parámetros de regularización $\lambda \in \{10, 7.9432, 6.3095, 5.0118, 3.9810, 3.1627, 2.5118, 1.9952, 1.5448, 1.2589, 1, 0.7943, 0.6309, 0.5011, 0.3981, 0.3161, 0.2511, 0.1995, 0.1584, 0.1258, 0.1\}$
4. Regresión logística con regularización Lasso. Con parámetros de regularización $\lambda \in \{10, 7.9432, 6.3095, 5.0118, 3.9810, 3.1627, 2.5118, 1.9952, 1.5448, 1.2589, 1, 0.7943, 0.6309, 0.5011, 0.3981, 0.3161, 0.2511, 0.1995, 0.1584, 0.1258, 0.1\}$
5. Máquinas de soporte vectorial:
   - Kernel lineal con
     - $\epsilon \in \{-1, -0.5, -0.1, 0, 0.1, 0.5, 1\}$
     - Costo $\in \{1, 5, 10, 25\}$
   - Kernel polinomial con
     - $\epsilon \in \{-1, -0.5, -0.1, 0, 0.1, 0.5, 1\}$
     - Costo $\in \{1, 5, 25\}$
     - Grados $\in \{2, 3, 4, 5\}$
     - Escala $\in \{0.1, 0.2, ..., 1.9, 2\}$
     - Offset $\in \{0.1, 0.2, ..., 0.9, 1\}$

- Kernel Gaussiano con
  - $\epsilon \in \{-1, -0.5, -0.1, 0, 0.1, 0.5, 1\}$
  - Costo $\in \{1, 5, 25\}$
  - $\sigma \in \{0, 0.1, 0.2, ..., 0.9, 1\}$
- Kernel tangente hiperbólico con
  - $\epsilon \in \{-1, -0.5, -0.1, 0, 0.1, 0.5, 1\}$
  - Costo $\in \{1, 5, 25\}$
  - Escala $\in \{0.1, 0.2, ..., 1.9, 2\}$
  - Offset $\in \{0.1, 0.2, ..., 0.9, 1\}$

6. Naive Bayes:
   - Laplace con hiperparámetros 0, 1, 3
7. Vecinos más cercanos: para este caso en particular los parámetros variarán dependiendo de cuál sea el k-pliegues que se este ocupando, pues para los primeros 4, los parámetros serán k $\in \{1, 10, 20, 30, 39, 59, 79, 99\}$, mientras que para el 5° pliegue, los parámetros ocupados serán k $\in \{1, 10, 20, 30, 40, 60, 80, 99\}$.
8. Bosques aleatorios con parámetros:
   - ntrees $\in \{100, 300, 500, 700, 1000\}$
   - mtrys $\in \{3, 4, 5\}$

Una vez que se especificaron los hiperparámetros para cada una de la técnicas, se comparará entre sí el desempeño y predicción de los mejores modelos para cada técnica, es decir, se comparará la mejor regresión logística contra la mejor Máquina de Soporte Vectorial; contra el mejor Random Forest, etc. Esta comparación y predicción se realizará sobre el conjunto de datos completo. Es decir, se ajustarán de nuevo cada uno de los modelos en un esquema de 80 % datos de entranamiento y 20 % de datos de prueba. En otras palabras, se tendrán los 8 mejores modelos (uno por cada técnica).

Una vez que se cuenta con estos 8 mejores modelos, se compararán dos a dos a partir de las pruebas no-paramétricas estudiadas en el Capítulo 3. Es decir, se llevarán a cabo 28 comparaciones con diferentes pruebas.

Posteriormente, se usarán la matriz de confusión y sus derivados, así como la curva ROC para obtener el mejor modelo globalmente.

5.3. Resultados de las pruebas de independencia

Se aplicaron las pruebas previamente vistas, en el capítulo 2, a las tres posibles combinaciones de los conjuntos de datos, de tal manera que la comparación fuera uno a uno, las combinaciones fueron IPC con INMEX, IPC con S&P 500 e INMEX con S&P 500.

La hipótesis nula que se aplicó para estas pruebas fue que los conjuntos de datos derivados de los movimientos de los índices accionarios del IPC, Inmex y S&P 500 son independientes uno a uno.

Los resultados de estas pruebas son los siguientes

Resultados de las pruebas de independencia	
Prueba realizada	p-valor
Ji-cuadrada IPC-INMEX	0
Ji-cuadrada IPC-SP500	8.0164e-60
Ji-cuadrada INMEX-SP500	1.5963e-56
Pearson IPC-INMEX	0
Pearson IPC-SP500	8.1616e-64
Pearson INMEX-SP500	5.7535e-60

Tomando en cuenta los resultados arrojados por la prueba de la Ji-cuadrada (χ^2), así como la prueba de la Correlación de Pearson, a través del p-valor, se puede rechazar la hipótesis nula de que los conjuntos de datos de los índices son independientes uno a uno, dado que todos los valores se encuentran por debajo del valor establecido para $\alpha_{0.95}$.

Este resultado es fácilmente interpretable cuando se habla de los indices IPC e Inmex, pues la composición de éstos llega a ser bastante similar, además de que ambos reflejan el comportamiento del mercado accionario dentro de una región geográfica muy específica, que es México. Por lo tanto, es natural pensar que si un hecho afecta a uno de los índices, muy probablemente el otro se comporte de la misma manera, con una mayor o menor medida.

Por otro lado, el resultado de los índices mexicanos con el índice estadounidense puede interpretarse como que existe una gran relación entre ambos mercados, pues si bien la fortaleza de los primeros no se compara con el segundo, aquellos hechos que lleguen a afectar al S&P 500 pueden repercutir en el mercado mexicano y por lo tanto, los índices se moverán en el mismo sentido.

Esto de igual manera ayudará a poder interpretar los resultados que se obtendrán más adelante.

5.4. Análisis descriptivo del conjunto de datos

El objetivo de esta sección es encontrar el mejor modelo de entre los mejores modelos de cada técnica, esto a través de las matrices de confusión y derivados de cada una de las ténicas; de curvas ROC para los 8 clasificadores; de la implementación de pruebas no paramétricas dos a dos, así como múltiples; y por último una revisión visual de las dispariedades entre los datos reales y los movimientos predichos por cada una de las técnicas.

Antes de poder revisar los resultados per-se de los modelos, primero habrá que revisar la información original y su comportamiento. En este caso, se mostrarán 3 tablas donde cada una indica el número de veces que los diversos índices subieron y bajaron a lo largo de los años, así como el promedio general en la ventana de tiempo analizada.

Para el caso del Índice de Precios y Cotizaciones, se puede observar que en promedio el número de alzas es ligeramente superior al número de bajadas. Pues tan solo un 1.34 % de los elementos es distinto, que representan 33 observaciones de las 2449 totales. Sin embargo se podría analizar, aunque no en este trabajo, el tamaño de las subidas o bajadas que ha tenido este índice para saber si realmente resulta rentable en el tiempo.

Índice de Precios y Cotizaciones					
Año	Incrementos	%	Decrementos	%	Total
2011	108	50.70	105	49.30	213
2012	136	54.84	112	45.16	248
2013	123	49.40	126	50.60	249
2014	124	49.40	127	50.60	251
2015	125	49.80	126	50.20	251
2016	138	54.76	114	45.24	252
2017	128	51.00	123	49.00	251
2018	116	50.20	125	49.80	251
2019	117	46.61	134	53.39	251
2020	116	50.00	116	50.00	232
Total	1241	50.67	1208	49.33	2449

Tabla con el número de incrementos y decrementos por año del IPC

La siguiente tabla muestra al Índice INMEX o Índice México el cuál tiene un comportamiento similar al IPC, aunque en este caso, el número de veces que baja el índice es ligeramente mayor comparado con el número de veces que el mismo sube. Sin embargo, las veces que el índice va a la baja es superior tan solo en 37 observaciones que representan el 1.5 % de las observaciones totales.

Es natural pensar que la diferencia de estos dos índices no debía ser muchas al estar basados ambos en la misma economía en desarrollo y además tener listadas acciones muy similares o iguales en ambos índices, aunque con ponderación distinta.

Índice México					
Año	Incrementos	%	Decrementos	%	Total
2011	102	47.66	112	52.34	214
2012	131	52.40	119	47.60	252
2013	123	48.81	129	51.19	252
2014	126	50.00	126	50.00	252
2015	124	49.21	128	50.79	252
2016	130	51.59	122	48.41	252
2017	127	50.60	124	49.40	251
2018	118	47.01	133	52.99	251
2019	119	47.22	133	52.78	252
2020	111	47.64	122	52.36	233
Total	1211	49.25	1248	50.75	2459

Tabla con el número de incrementos y decrementos por año del Inmex

A diferencia de los índices anteriores, el S&P 500, muestra una diferencia más marcada entre las veces que sube dicho índice a las veces que baja, siendo superiores el número de subidas, con al rededor de 235 días que representan un 9.56 % de las observaciones.

Esto podría deberse a que en este índice están listadas empresas con origen en una economía mucho más sólida y desarrollada, a diferencia de las empresas listadas en el IPC o en el INMEX.

Standar & Poor's 500					
Año	Incrementos	%	Decrementos	%	Total
2011	114	53.27	100	46.73	214
2012	132	52.80	118	47.20	250
2013	147	58.33	105	41.67	252
2014	144	57.14	108	42.86	252
2015	119	47.22	133	52.78	252
2016	131	51.98	121	48.02	252
2017	143	56.97	108	48.03	251
2018	132	52.59	119	47.41	251
2019	150	59.52	102	40.48	252
2020	135	57.94	98	42.06	233
Total	1347	54.78	1112	45.22	2459

Tabla con el número incrementos y decrementos por año del S&P 500

5.5. Ejecución de los modelos seleccionados

Una vez que se corrieron las diversas pruebas se obtienen los hiperparámetros con el mejor desempeño para cada uno de los distintos modelos, dependiendo del índice, a continuación se mostrará una tabla con dichos hiperparámetros.

Modelo	IPC	Inmex	SP500
R. Logística	NA	NA	NA
R. Probit	NA	NA	NA
R. Ridge	$\lambda = 0.1$	$\lambda = 0.1$	$\lambda = 0.1$
R. Lasso	$\lambda = 0.1$	$\lambda = 0.1$	$\lambda = 0.1$
MSV	Kernel = rbfdot $\epsilon = -1$ costo = 1 $\sigma = 0.3$	Kernel = rbfdot $\epsilon = -1$ costo = 1 $\sigma = 0.1$	Kernel = rbfdot $\epsilon = -1$ costo = 1 $\sigma = 0.1$
Naive-Bayes	Laplace $\in \{0, 1, 3\}$	Laplace $\in \{0, 1, 3\}$	Laplace $\in \{0, 1, 3\}$
Bosques Aleatorios	ntrees = 1000 mtrys = 5 sampz = 1184	ntrees = 100 mtrys = 3 sampz = 1188	ntrees = 100 mtrys = 3 sampz = 1287
K-Vecinos Cercanos	vecino = 99	vecino = 99	vecino = 99

Tabla con los hiperparámetros finales de los diversos modelos utilizados

5.6. Comparación de resultados

5.6.1. Resultados del Índice de Precios y Cotizaciones

Lo siguiente a revisar, posterior al análisis de los índices, son los indicadores utilizados, es decir, Media Móvil Simple, Indicador de Momentum, Indicador Estocástico K, Indicador Estocástico D, Indice de Fuerza Relativa (RSI), la Media Móvil de Convergencia/Divergencia (MACD) y el Oscilador de Acumulación/Distribución.

La finalidad de este anális es poder observar el comportamiento de los índicadores contra si mismos, en forma de distribución, así como contra sus pares y ver la correlación que se guarda entre los mismos, de tal manera de poder detectar elementos que estén fuertemente correlacionados y eliminarlos.

Por último, se añadió una nube de puntos con la finalidad de detectar de manera visual y rápida la separación entre las ocasiones en las que sube el índice y las que baja, así como el desempeño del indicador en cada ocasión.

Los elementos que compondrán las siguientes gráficas son:

1. Diagonal: en ella se encuentran las distribuciones de subidas y bajadas hechas por cada uno de los indicadores.

2. Triángulo superior: se encuentra el nivel de correlación entre los diversos indicadores.

3. Triángulo inferior: muestra la nube de puntos con las distribuciones de las subidas y bajadas de los mismos, por relación entre indicadores.

En la gráfica anterior se puede observar lo siguiente:

1. El comportamiento de la media móvil simple asemeja a una distribución normal y es de esperarse pues este indicador únicamente replica el comportamiento promedio del precio del índice a lo largo del tiempo. Este comportamiento aleatorio será similar en los otros dos índices.

2. En cuanto a los mejores comportamientos de separación entre las subidas y bajadas, cabe resaltar el Oscilador de Acumulación/Distribución (Osc AD), el cuál separa de manera bastante efectiva ambas clases. Lo que hace pensar que este indicador sea fuertemente usado por los modelos para lograr predecir las subidas y bajadas en los conjuntos de entrenamiento.

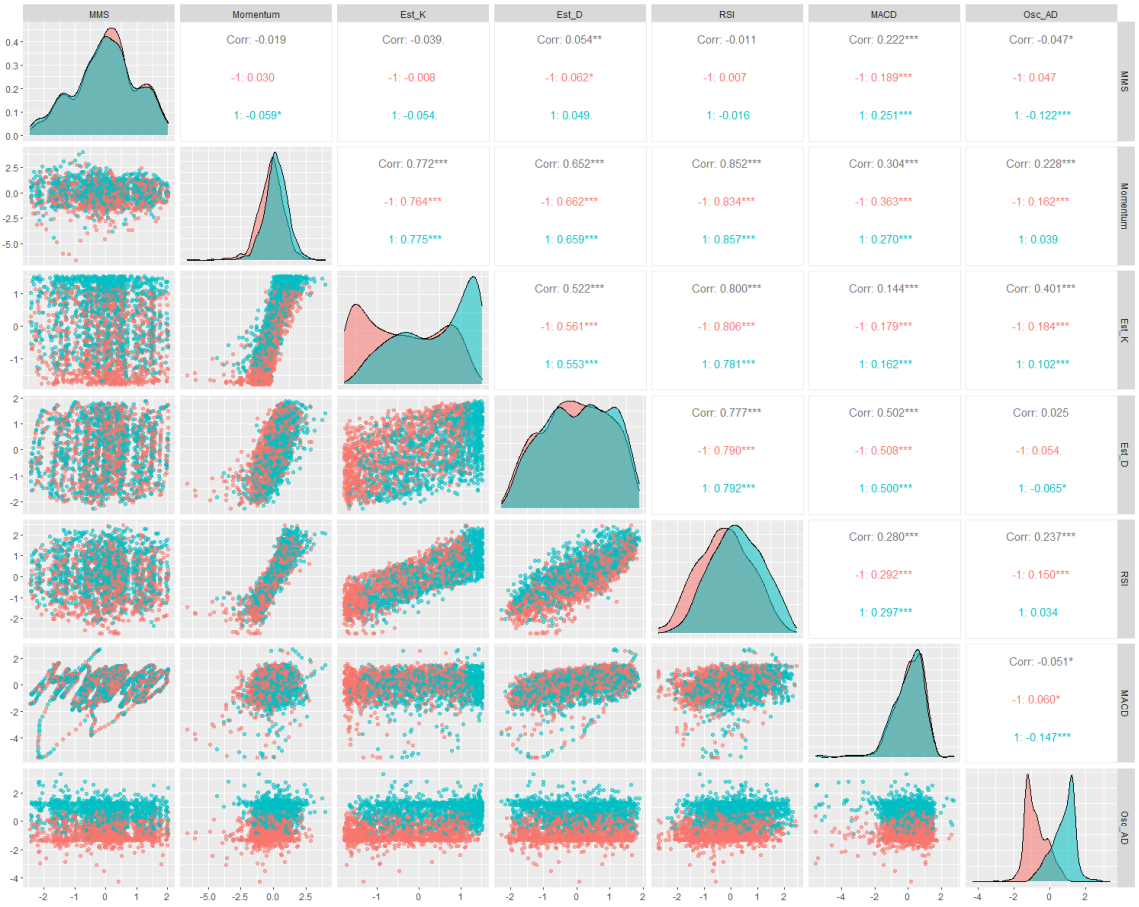


Figura 5.2: *GGPairs de los índicadores aplicados sobre el IPC*

3. Entre los indicadores más correlacionados se encuentran el RSI y el Indicador de Momentum; el RSI y el Oscilador Estocástico K; y el RSI y el Oscilador Estocástico D. En cierto punto, esto llega a tener sentido pues estos elementos basan su análisis en un mismo componente que es el impulso o la velocidad con la que cambian los precios de un momento a otro.

Para el análisis del desempeño de los modelos, primero se mostrará una gráfica con el conjunto de resultados de las matrices de confusión arrojadas por cada uno de los modelos aplicados al conjunto de datos de prueba. Hay que recordar que para llegar a este punto, primero se debieron de entrenar los modelos con el conjunto de prueba, y los datos y parámetros obtenidos se encuentran descritos en la sección de 4 de este capítulo, Ejecución de los modelos seleccionados.

Adicionalmente, dado que hay más de una forma de medir el desempeño de un modelo,

se analizarán algunas medidas tales como la precisión, la sensibilidad, la especificidad, etc. Como tercer paso de análisis se mostrarán las curvas ROC, las áreas que éstas abarcan y cada uno de sus cutoff. Y por último, se decidió hacer un análisis no paramétrico sobre el desempeño de los modelos con la finalidad de poder decidir de manera un poco más rigurosa, estadísticamente hablando, si realmente existe alguna diferencia importante entre alguno de estos modelos.

Al finalizar estos análisis se mostrarán una serie gráficas las cuales representan la diferencia entre la serie de tiempo con sus respectivas subidas y bajadas, y las diferencias que tuvieron los modelos con las subidas y bajadas reales.

5.6.1.1. Variables de decisión del bosque aleatorio IPC

En esta sección se analizará brevemente cuáles fueron las variables que más tuvieron peso en la toma de decisión del bosque aleatorio para el índice IPC. Para ello se tomará en cuenta el resultado mostrado en las siguientes dos gráficas, donde en la primera de lado izquierdo se puede observar qué tanto decaería la predicción del modelo de Bosques Aleatorios si removiecemos alguno de los indicadores, mientras que en el gráfico de lado derecho se muestra la pureza de los nodos al final de los árboles de decisión si se removiece alguno de estos indicadores.

Se puede ver claramente como en la gráfica de "MeanDecreaseAccuracy" del lado izquierdo, el predictor con un mayor peso es el Oscilador de Acumulación/Distribución, al estar presente en más de 300 resultados. En segundo lugar y bastante por debajo se encuentra el Oscilador Estocástico K.

Por otro lado, en la gráfica de pureza de nodos o "MeanDecreaseGini", se puede ver que de igual manera en primer lugar se encuentra el Oscilador de Acumulación/Distribución, seguido del Oscilador Estocástico K.

Este resultado va de la mano con lo visto en la gráfica de GGPairs, donde es claro que un buen indicador para separar a los grupos de subida y bajada es el Oscilador A/D.

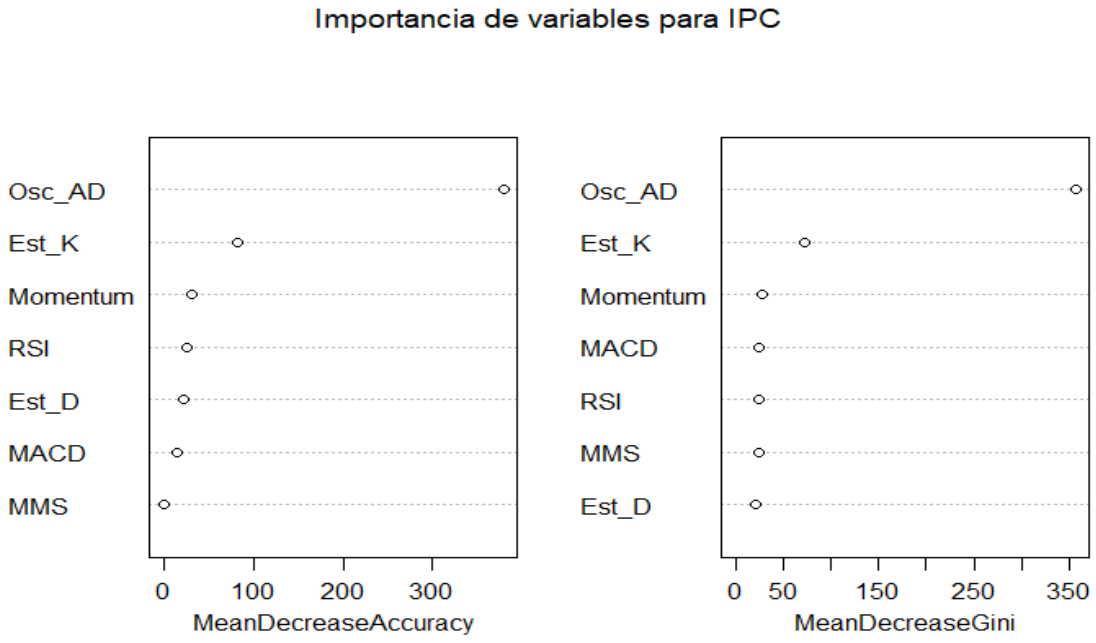


Figura 5.3: *Gráfica con las variables de mayor importancia para el Bosque Aleatorio del IPC*

Adicionlamente, a continuación se muestra una tabla con los valores numéricos del indicador de Gini y el número real donde los predictores son usados en el Bosque Aleatorio.

Uso de variables por Bosque Aleatorio		
Indicador	MediaCaidaGini	Predictores usados
Oscilador AD	356.2006	17281
Estocástico K	73.4758	14577
Momentum	27.5773	10982
MACD	24.9722	13024
RSI	24.4194	10600
MMS	23.5738	12284
Estocástico D	21.2183	10671

5.6.1.2. Matriz de Confusión IPC

Comenzando con las matrices de confusión, se puede observar como los modelos que tienen un mejor desempeño son el bosque aleatorio, así como la máquina de soporte vectorial,

al tener una menor cantidad de observaciones mal clasificadas, con 52 y 64 respectivamente.

Otros modelos que obtuvieron un buen desempeño, son las regresiones, pues sus observaciones mal clasificadas están entre los 67 y los 69 elementos.

Por último se encuentran los modelos de K-vecinos cercanos y Naive-Bayes, con 72 y 76 elementos mal clasificados, respectivamente. En este caso se puede observar, como el modelo de Naive-Bayes fue el que peor desempeño tuvo a simple vista con las matrices de confusión.

Figura 5.4: *Conjunto de las Matrices de Confusión de los modelos aplicados al IPC*

5.6.1.3. Métricas de clasificadores discretos IPC

En la siguiente tabla se mostrarán algunas métricas obtenidas a partir de la matriz de confusión, también conocidos como clasificadores discretos, los cuáles están ordenados de

forma descendente por la medida "Accuracy" o exactitud de clasificación. Adicionalmente también se tiene la sensibilidad, la especificidad, la precisión, el recall y la medida F1; medidas de las cuales se hace mención en el capítulo 2.

Los resultados para el IPC, muestran que para todas las métricas, el mejor desempeño lo tuvieron tanto el método de bosques aleatorios (random forest) y las máquinas de soporte vectorial.

Posteriormente, las regresiones tanto Logit, Probit, Ridge y Lasso presentan una precisión muy similar; y por último se tienen tanto al modelo de K-vecinos cercanos (KNN), y al modelo de Naive-Bayes.

En cuanto al resto de las medidas, no se puede declarar un modelo superior a los otros para todas las categorías, sin embargo, si se puede observa como el modelo Naive-Bayes resulta ser inferior en todas ellas, comparandolo con sus pares.

Modelo	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
Random Forest	0.8951613	0.9098361	0.8809524	0.8809524	0.9098361	0.8951613
MSV	0.8709677	0.8760000	0.8658537	0.8690476	0.8760000	0.8725100
Probit	0.8649194	0.8714859	0.8582996	0.8611111	0.8714859	0.8662675
Ridge	0.8649194	0.8685259	0.8612245	0.8650794	0.8685259	0.8667992
Logit	0.8629032	0.8709677	0.8548387	0.8571429	0.8709677	0.8640000
Lasso	0.8608871	0.8616601	0.8600823	0.8650794	0.8616601	0.8633663
KNN	0.8548387	0.8600000	0.8495935	0.8531746	0.8600000	0.8565737
Naive-Bayes	0.8467742	0.8548387	0.8387097	0.8412698	0.8548387	0.8480000

Figura 5.5: *Métricas derivadas de las matrices de confusión para el IPC*

5.6.1.4. Curvas ROC IPC

En la siguiente tabla se muestran los valores del área que cubren las diversas curvas ROC, así como el cutoff de cada una de ellas. La tabla esta ordenada de manera descendente por el área de las curvas.

En primer lugar se encuentran las regresiones Ridge, Probit y Logit, con un área muy cercana entre si, superior a los 0.952. A continuación, y ligeramente por debajo, se encuentra la regresión Lasso con 0.951. A partir de este punto el área bajo la curva se reduce considerablemente a niveles inferiores a los 0.9.

En quinto y sexto lugar en cuantoa a tamaño del área bajo la curva se encuentra el método de bosque aleatorio y máquinas de soporte vectorial, con un área respectiva de 0.89 y 0.87.

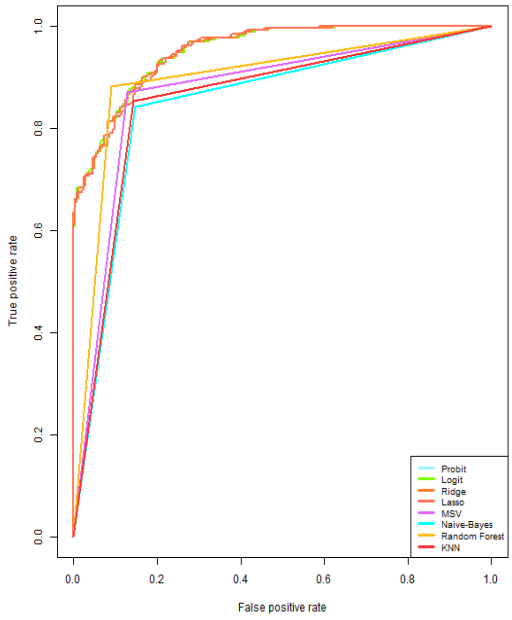
En último lugar se encuentra nuevamente el modelo de Naive-Bayes donde posé un área de 0.84, que si bien resulta ser aceptable, comparada con sus pares es la menor de ellas. Además, este resultado es consistente con lo visto anteriormente en la tabla de medidas de las matrices de confusión

En cuanto al umbral de decisión o cuttoff, se puede ver que los modelos de bosques aleatorios, máquinas de soporte vectorial, k-vecinos cercanos y Naive-Bayes, presentan un cutoff exacto de 1, mientras que todas las regresiones muestran un cutoff menor a uno. Esto se podrá ver de manera más clara a continuación en el gráfico correspondiente.

En el gráfico es claro, como el cutoff de los modelos influye directamente en el comportamiento de las curvas, al hacer un solo corte para los modelos supervisados y múltiples saltos para las regresiones, regularizadas y no regularizadas.

Modelo	Area	Cutoff
Ridge	0.9522508	0.4756095
Probit	0.9522021	0.4933099
Logit	0.9521533	0.4917787
Lasso	0.9513564	0.4642747
Random Forest	0.8953942	1.0000000
MSV	0.8709992	1.0000000
KNN	0.8548660	1.0000000
Naive-Bayes	0.8468644	1.0000000

(a) *Área Curva Roc y Cutoff*



(b) *Curvas Roc*

Figura 5.6: *Conjunto de las Curvas Roc generada por los diferentes modelos*

5.6.1.5. Errores de predicción de las series de tiempo IPC

En las siguientes gráficas se hace una representación visual de en qué momentos de la serie de tiempo, los distintos modelos fallan en la predicción del movimiento de los índices. Dependiendo del modelo aumentará o disminuirá el número de puntos azules (las diferencias de tendencia) que aparecerán sobre la serie de tiempo en color naranja.

Estas gráficas se encuentran dibujadas en un fondo separado de manera mensual, para que sea más fácil la ubicación de los momentos en los cuáles hubo discordancia entre el movimiento real y el movimiento pronosticado.

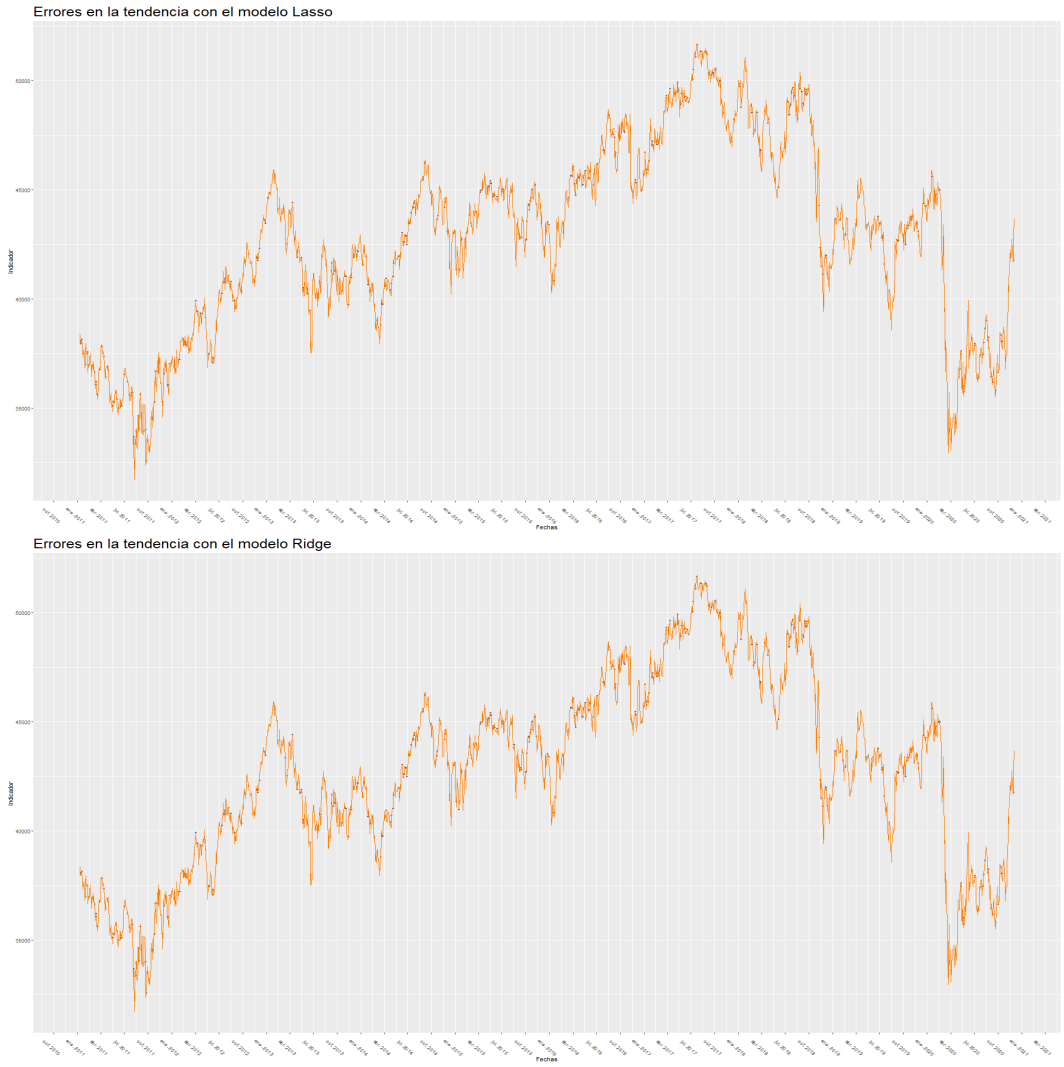


Figura 5.7: *Series de tiempo con diferencias en la predicción por parte de los modelos*

En las dos gráficas anteriores, tanto en la regresión Ridge, como Lasso, se puede apreciar que las predicciones presentan problemas de asignación cuando se hace un cambio de tendencia, principalmente de un movimiento al alza con uno a la baja. Sin embargo, se puede apreciar que la concentración de los puntos y las ubicaciones son muy similarles entre si a lo largo de la serie del tiempo.

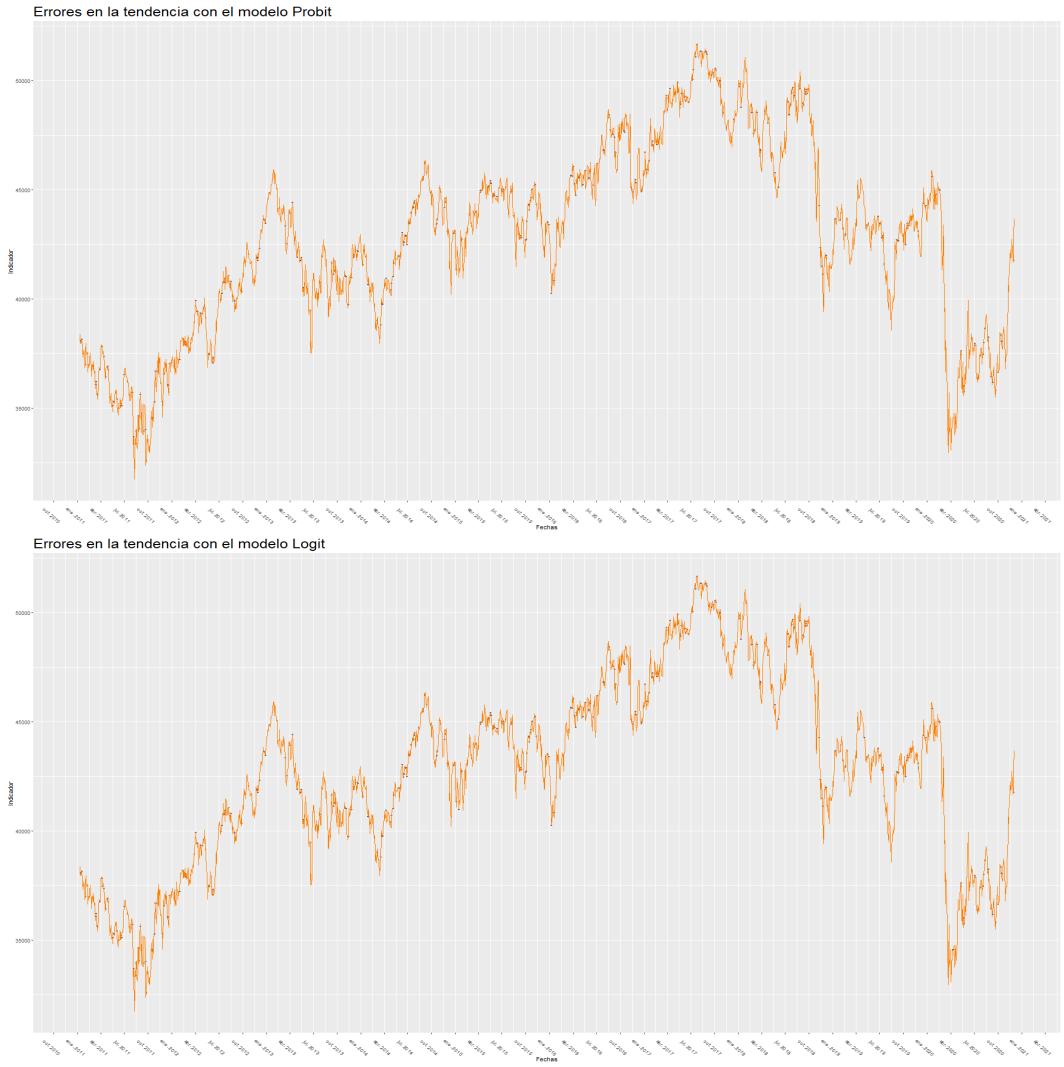


Figura 5.8: *Series de tiempo con diferencias en la predicción por parte de los modelos*

Comparando las gráficas de diferencias entre las predicciones de los modelos probit y logit, estas son similares entre si, y a diferencia de las regresiones Ridge y Lasso, aquí si llegan a presentarse diferencias entre las predicciones y los movimientos reales mientras se tiene algún movimiento, ya sea completamente a la baja o al alza; además claro de presentar una mayor dificultad al predecir los movimientos en periodos de mayor volatilidad.

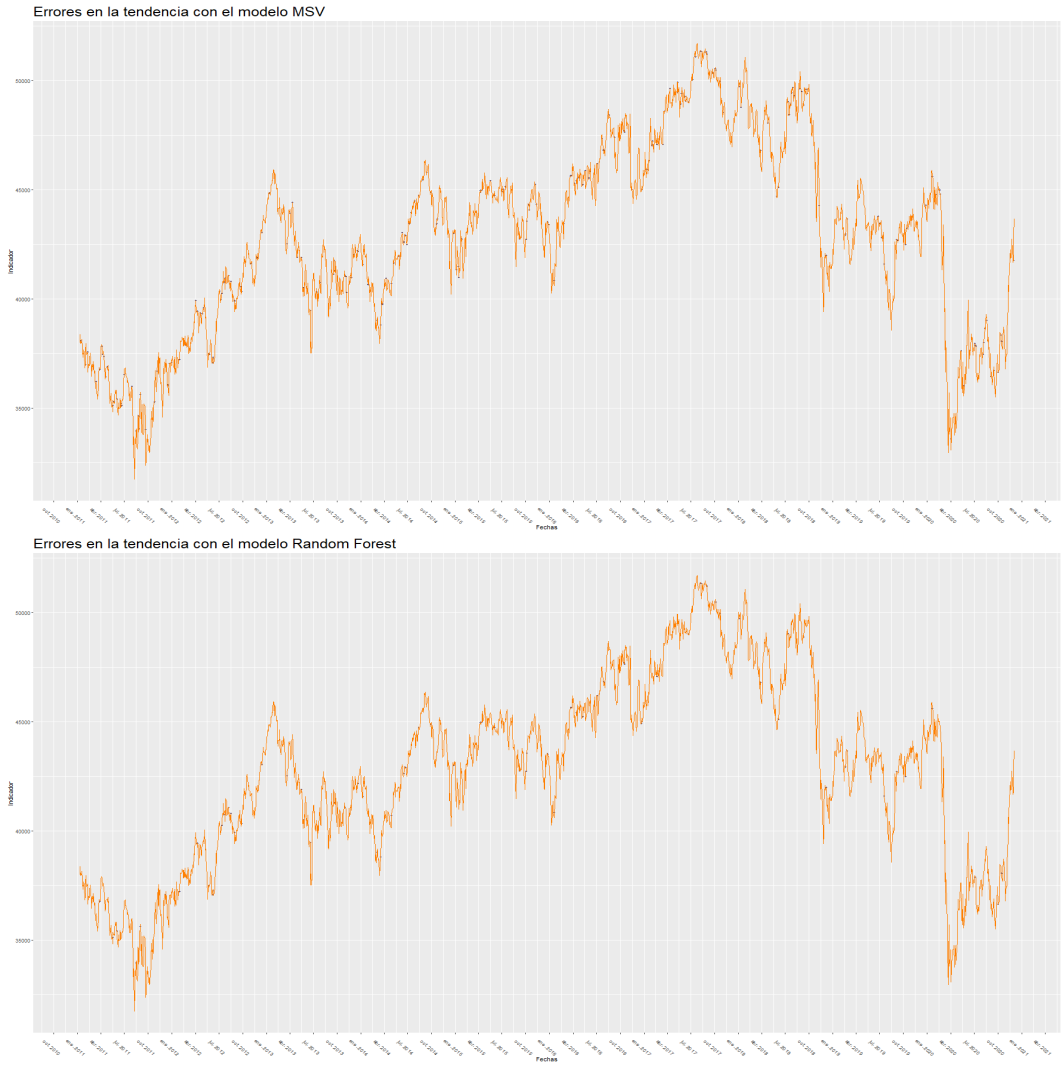


Figura 5.9: *Series de tiempo con diferencias en la predicción por parte de los modelos*

Con respecto a las gráficas anteriores, se puede observar que el modelo de bosques aleatorios presenta un menor número de elementos incorrectos entre la predicción y los movimientos reales, siendo estos más claros cuando se hace un cambio de tendencia en el movimiento de los precios. Por otra parte, se puede observar como las máquinas de soporte vectorial presentan una mayor dificultad al predecir los cambios bruscos de movimientos entre periodos cortos de tiempo, en particular entre enero y diciembre del 2015, o entre febrero del 2016 al 2017.

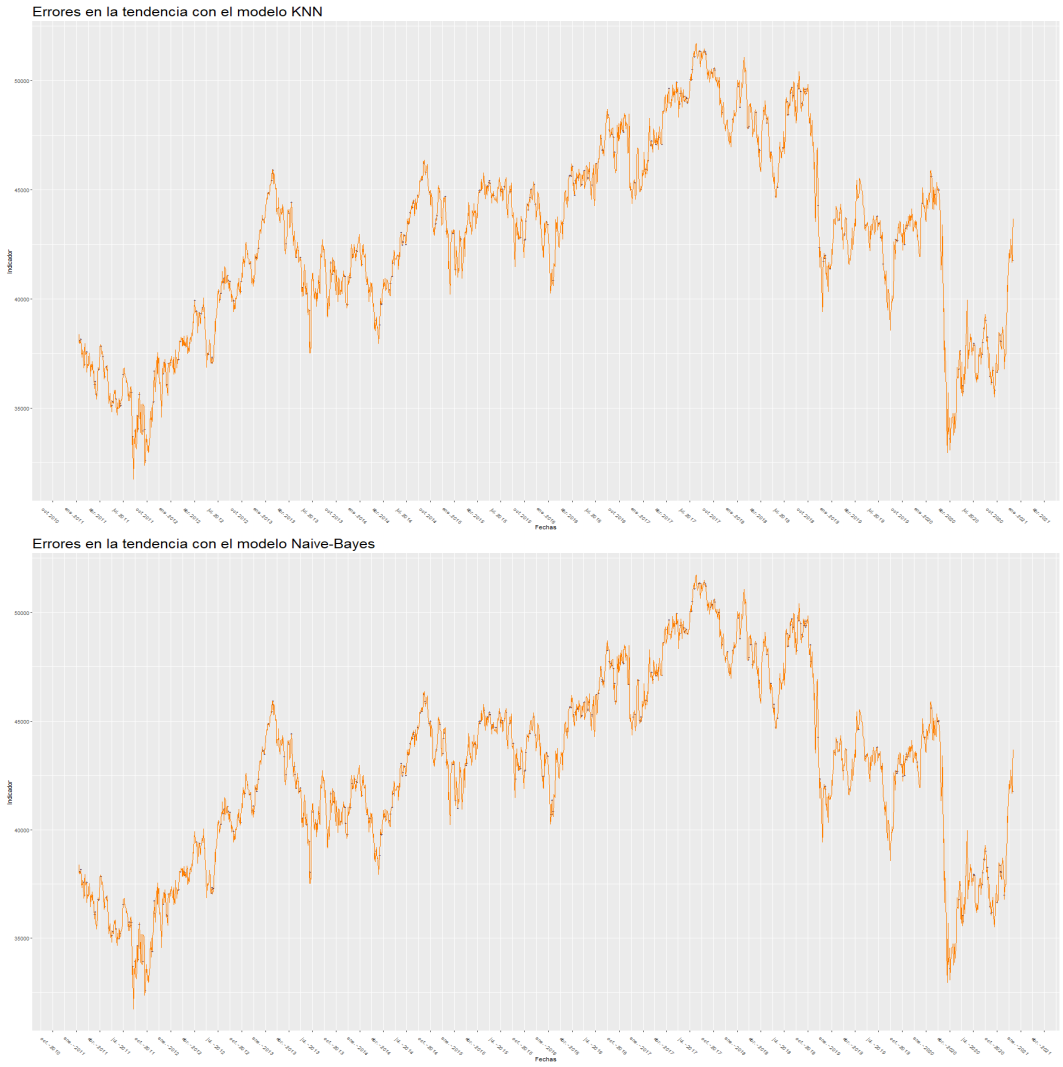


Figura 5.10: *Series de tiempo con diferencias en la predicción por parte de los modelos*

Por último, las gráficas de KNN y Naive-Bayes presentan un número más elevado de movimientos mal clasificados, teniendo errores tanto en cambios de tendencia, así como en tendencias sostenidas, por ejemplo en el periodo de julio del 2012 a febrero del siguiente año. Sin embargo, entre estos dos modelos, se puede apreciar como Naive-Bayes presenta una mayor dificultad para poder hacer una predicción certera en micromovimientos del mercado o en épocas de gran volatilidad.

Recordar que para todas estas gráficas el modelo se aplico sobre el conjunto completo de datos a diferencia de las matrices de confusión donde únicamente son los resultados de

los conjuntos de prueba.

5.6.1.6. Pruebas no paramétricas entre los diversos modelos para IPC

En las siguientes tablas se muestran las diversas comparaciones uno a uno de los resultados obtenidos por los diferentes modelos de pronóstico, donde en color verde se mostrarán los elementos que no muestran una diferencia significativa de acuerdo al p-value arrojado por la prueba correspondiente. En color rojo se mostrarán los elementos cuya diferencia es significativa y el p-valor es muy pequeño. Cabe mencionar que estas pruebas se realizaron con un nivel de significancia del 0.95 %.

En en la tabla correspondiente a la prueba de Signos se puede ver que todos los elementos están en rojo, sin embargo, esto se puede atribuir a que dicha prueba no esta preparada del todo para clasificar elementos binarios, es decir, cuyas clases únicamente posean valores 0 ó 1. Este comportamiento se repetirá en los siguientes índices.

Por otro lado, la prueba de Wilcoxon presenta todos sus valores en verde, es decir, bajo esta prueba no se encuentran diferencias significativas entre el desempeño de los diferentes modelos ejecutados para el IPC.

Prueba Signos	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0	0	0	0	0	0	0
Naive-Bayes	0	0	0	0	0	0	NA
RandForest	0	0	0	0	0	NA	NA
MSV	0	0	0	0	NA	NA	NA
Ridge	0	0	0	NA	NA	NA	NA
Lasso	0	0	NA	NA	NA	NA	NA
Probit	0	NA	NA	NA	NA	NA	NA

Prueba Wilcoxon	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.8991	0.9495	0.8491	0.9495	1.0000	0.7034	0.8991
Naive-Bayes	1.0000	0.9495	0.7511	0.8491	0.8991	0.7997	NA
RandForest	0.7997	0.7511	0.5679	0.6569	0.7034	NA	NA
MSV	0.8991	0.9495	0.8491	0.9495	NA	NA	NA
Ridge	0.8491	0.8991	0.8991	NA	NA	NA	NA
Lasso	0.7511	0.7997	NA	NA	NA	NA	NA
Probit	0.9495	NA	NA	NA	NA	NA	NA

Figura 5.11: *Prueba no paramétrica de Signos y Wilcoxon aplicado a los diferentes modelos de predicción*

Prueba Friedman	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.6698	0.8273	0.4913	0.8185	1.0000	0.3428	0.715
Naive-Bayes	1.0000	0.8658	0.3692	0.6015	0.7855	0.6171	NA
RandForest	0.5465	0.4458	0.1699	0.2743	0.2207	NA	NA
MSV	0.7237	0.8618	0.6015	0.8618	NA	NA	NA
Ridge	0.0833	0.1573	0.1573	NA	NA	NA	NA
Lasso	0.0253	0.0455	NA	NA	NA	NA	NA
Probit	0.3173	NA	NA	NA	NA	NA	NA

Prueba Davenport	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	H0	H0	H0	H0	H0	H0	H0
Naive-Bayes	H0	H0	H0	H0	H0	H0	NA
RandForest	H0	H0	H0	H0	H0	NA	NA
MSV	H0	H0	H0	H0	NA	NA	NA
Ridge	H0	H0	H0	NA	NA	NA	NA
Lasso	H1	H1	NA	NA	NA	NA	NA
Probit	H0	NA	NA	NA	NA	NA	NA

Figura 5.12: *Prueba no paramétrica de Friedman y Davenport aplicado a los diferentes modelos de predicción*

Bajo la prueba no paramétrica de Friedman, se puede observar como el desempeño de la regresión Ridge presenta una diferencia significativa en relación a la regresión Logit y Probit, sin embargo, no lo hace para el resto de las pruebas. Adicionalmente, en la prueba de Iman Davenport también se puede apreciar commo se repite el mismo comportamiento que en la prueba de Friedman.

Rangos Alineados	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.8791	0.9393	0.8190	0.9392	1.0000	0.6543	0.8801
Naive-Bayes	1.0000	0.9402	0.7064	0.8215	0.8831	0.7710	NA
RandForest	0.7662	0.7099	0.5031	0.6017	0.6489	NA	NA
MSV	0.8804	0.9401	0.8215	0.9401	NA	NA	NA
Ridge	0.8161	0.8766	0.8766	NA	NA	NA	NA
Lasso	0.6988	0.7567	NA	NA	NA	NA	NA
Probit	0.9381	NA	NA	NA	NA	NA	NA

Prueba Quade	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.6703	0.8275	0.4918	0.8188	1.0000	0.3433	0.7154
Naive-Bayes	1.0000	0.8660	0.3697	0.6020	0.7858	0.6176	NA
RandForest	0.5470	0.4463	0.1702	0.2747	0.2210	NA	NA
MSV	0.7241	0.8620	0.6020	0.8620	NA	NA	NA
Ridge	0.0833	0.1575	0.1575	NA	NA	NA	NA
Lasso	0.0252	0.0454	NA	NA	NA	NA	NA
Probit	0.3178	NA	NA	NA	NA	NA	NA

Figura 5.13: *Prueba no paramétrica de Rangos Alineados y Quade aplicado a los diferentes modelos de predicción*

Tomando en cuenta la prueba de Rangos Alienados, no se encuentra ninguna diferencia significativa entre el desempeño de nignuno de los modelos, empero, observando la prueba de Quade, se puede observar como la regresión Ridge vuelve a presentar una diferencia significativa con las regresiones Logit y Probit.

Prueba	P_value
Friedman	0.8319
RangosAlin	0.9944
Quade	0.8322
Davenport	H0

Figura 5.14: *Pruebas no paramétricas múltiples aplicadas a los diferentes modelos de predicción*

Por último, se tienen las pruebas múltiples aplicadas a todos los resultados de los modelos utilziados. Las pruebas utilizadas que tienen esta capacidadd de análisis múltiple fueron la prueba de Friedman, Rangos Alineados, Quade e Iman Davenport, donde los resultados de estas cuatro pruebas muestran que no existe alguna diferencia significativa entre los elementos analizados.

5.6.2. Resultados del Índice México

Continuando con el análisis de los resultados, el siguiente índice a revisar es el Índice México, que en términos generales tiene un comportamiento muy similar al IPC, pues su composición y su diversificación regional tiene varias similitudes.

El primer punto a revisar de los resultados obtenidos es el gráfico de pares, el cuál tiene la misma composición que el gráfico de IPC visto previamente, es decir, contiene una distribución de subidas y abajadas en la diagonal, un nivel de correlación en el triángulo superior y una nube de puntos en el triángulo inferior.

En el siguiente gráfico podemos observar:

1. Como ya se había mencionado, el comportamiento de esta gráfica es muy similar tanto en la distribución de los Indicadores, como en sus niveles de correlación entre si y las nubes de puntos con las dos clases que se observan, con el Índice de Precios y Cotizaciones.

2. El indicador que tiene un comportamiento más similar de su distribución a una distribución normal, vuelve a ser la media móvil simple, por el mismo comportamiento de los puntos que la conforman.

3. En el caso del Índice México, se vuelve a repetir el comportamiento del Oscilador de Acumulación/Distribución, al ser el indicador que mejor separa las clases de subida y de bajada en la nube de puntos. Así mismo, la distribución que presentan estos puntos es visiblemente contraria en la diagonal de la gráfica.

4. Nuevamente entre los indicadores más correlacionados se encuentran el RSI y el Indicador de Momentum; el RSI con el Oscilador Estocástico K y D; así como el Momentum y el Oscilador K. Hay que recordar que estos tres indicadores basan su análisis en la velocidad con la que cambian los precios de un momento a otro del tiempo.

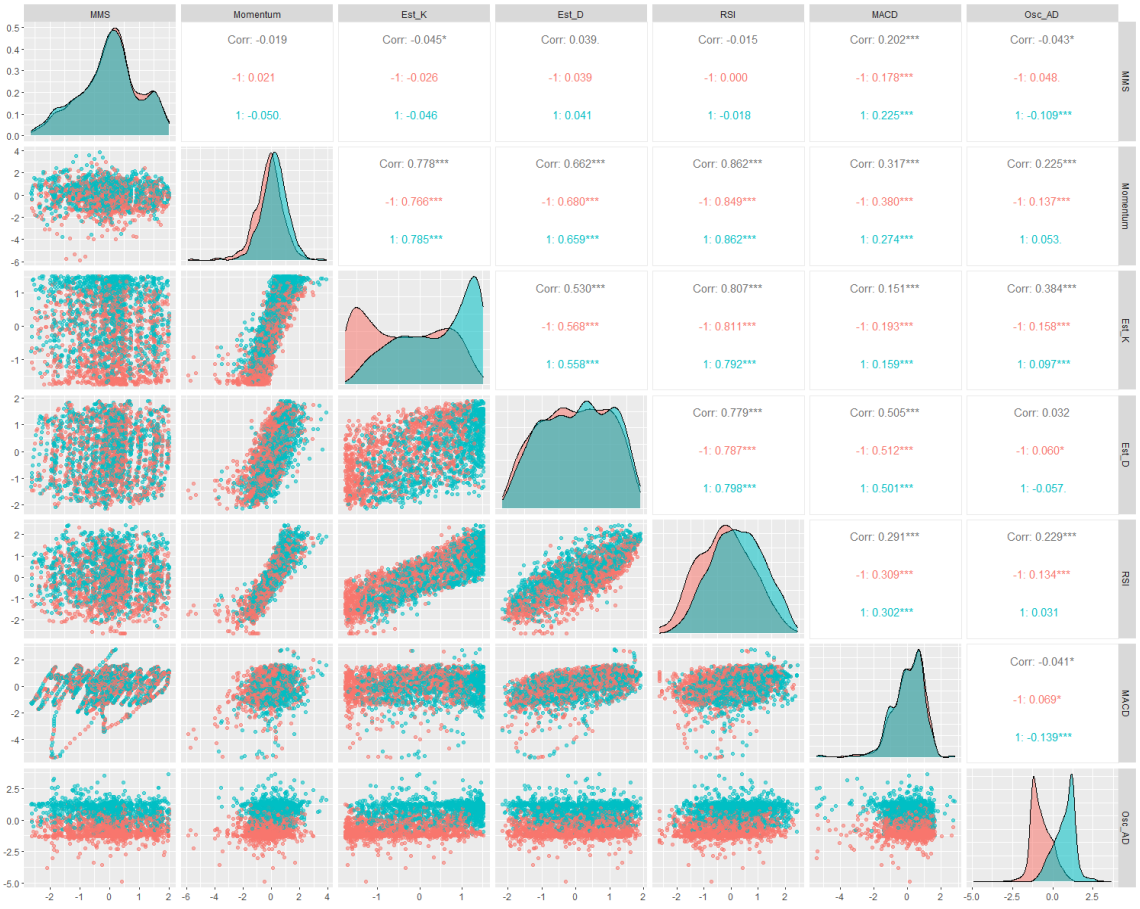


Figura 5.15: *GGPairs de los índicadores aplicados sobre el INMEX*

5.6.2.1. Variables de decisión del bosque aleatorio INMEX

En esta sección se analizará brevemente cuáles fueron las variables que más tuvieron peso en la toma de decisión del bosque aleatorio para el índice INMEX. Para ello se tomará en cuenta el resultado mostrado en las siguientes dos gráficas, donde en la primera de lado izquierdo se puede observar qué tanto decaería la predicción del modelo de Bosques Aleatorios si removiecemos alguno de los indicadores, mientras que en el gráfico de lado derecho se muestra la pureza de los nodos al final de los árboles de decisión si se removiece alguno de estos indicadores.

En la gráfica de "MeanDecreaseAccuracy" se puede observar como nuevamente para el conjunto de datos del índice INMEX, el Oscilador de Acumulación/Distribución es el que se presenta como uno de los predictores más importantes, seguido del Oscilador Estocástico K.

De igual manera, en la gráfica de "MeanDecreaseGini" la cual mide la pureza de los nodos, se puede ver que los indicadores que repiten posición al igual que en el IPC, son el Oscilador A/D y el Oscilador Estocástico K.

Estos resultados van en concordancia nuevamente con el GGPairs del INMEX que se vió previamente.

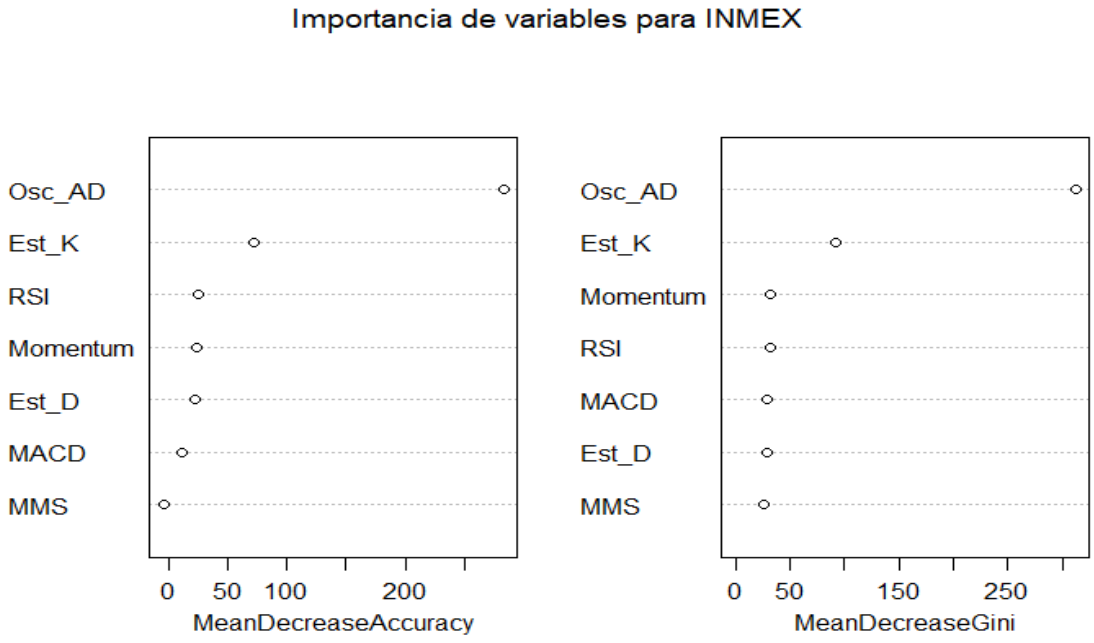


Figura 5.16: *Gráfica con las variables de mayor importancia para el Bosque Aleatorio del INMEX*

En cuanto a la siguiente tabal, se puede ver numéricamente que el indicador que más fue utilizado en el Bosque Aleatorio para el Índice México, es el Oscilador de Acumulación/Distribución y el Oscilador Estocástico K. A diferencia del IPC, en este caso la diferencia entre el uso de un indicador y otro llega a ser ligeramente menor, sin embargo, siguen siendo importantes.

Uso de variables por Bosque Aleatorio		
Indicador	MediaCaidaGini	Predictores usados
Oscilador AD	312.5501	15601
Estocástico K	90.8061	13042
RSI	33.0741	10372
Momentum	32.9847	10322
Estocástico D	29.2786	10124
MACD	29.1352	10616
MMS	26.1821	10200

5.6.2.2. Matriz de confusión INMEX

El siguiente resultado a revisar es la matriz de confusión, pues es una forma más clara de observar y comparar el desempeño de los modelos ejecutados.

En la siguiente comparación se puede observar como, los elementos con un menor número de clasificaciones mal realizadas son tanto la máquinas de soporte vectorial como la regresión Lasso, con 66 y 67 elementos respectivamente. Esto contrasta un poco con los elementos del IPC donde el bosque aleatorio es aquel con un mejor resultado.

Adicionalmente, las regresiones tales como Ridge y Probit presentan un buen desempeño a la hora de clasificar elementos, con pocos elementos mal clasificados. con 69 y 70 elementos respectivamente. En contraste, y similar al resultado mostrado en el IPC, el modelo de Naive-Bayes es el que muestra un peor comportamiento a la hora de clasificar, con 81 elementos mal clasificados.

Figura 5.17: *Conjunto de las Matrices de Confusión de los modelos aplicados al IPC*

5.6.2.3. Métricas de clasificadores discretos INMEX

El tercer elemento a considerar en el análisis de resultados para el Índice México son las métricas obtenidas a partir de las matrices de confusión vistas anteriormente, también conocidas como clasificadores discretos. La tabla que se muestra a continuación está ordenada de forma descentende por la medida "Accuracy" o exactitud de clasificación. Para mayor detalle del significado de las métricas utilizadas consultar el capítulo 2.

Los resultados para el INMEX, muestran que los modelos con un mejor desempeño son las máquinas de soporte vectorial para todas las métricas, seguido muy de cerca por la regresión Lasso y Ridge, posteriormente y no muy por debajo, los demás modelos; donde el peor comportamiento lo muestra el modelo de Naive-Bayes.

En relación a las demás métricas, aunque en general se muestra un comportamiento decendiente de los modelos, similar a la métrica de "accuracy", se tienen algunas diferencias como en el modelo de KNN, donde si bien la precisión es menor que otros modelos, medidas como la especificidad o el recall es superior a otros modelos.

Modelo	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
MSV	0.8674699	0.8612245	0.8735178	0.8683128	0.8612245	0.8647541
Lasso	0.8654618	0.8606557	0.8700787	0.8641975	0.8606557	0.8624230
Ridge	0.8614458	0.8565574	0.8661417	0.8600823	0.8565574	0.8583162
Probit	0.8594378	0.8530612	0.8656126	0.8600823	0.8530612	0.8565574
Random Forest	0.8594378	0.8559671	0.8627451	0.8559671	0.8559671	0.8559671
Logit	0.8574297	0.8495935	0.8650794	0.8600823	0.8495935	0.8548057
KNN	0.8574297	0.8524590	0.8622047	0.8559671	0.8524590	0.8542094
Naive-Bayes	0.8373494	0.8240000	0.8508065	0.8477366	0.8240000	0.8356998

Figura 5.18: *Métricas derivadas de las matrices de confusión para el INMEX*

5.6.2.4. Curvas ROC INMEX

En el gráfico de la curva ROC, se puede apreciar como las regresiones tanto regularizadas como no regularizadas muestran una forma más "curva", es decir, el movimiento del gráfico es más suave que la gráfica generada por los otros cuatro modelos.

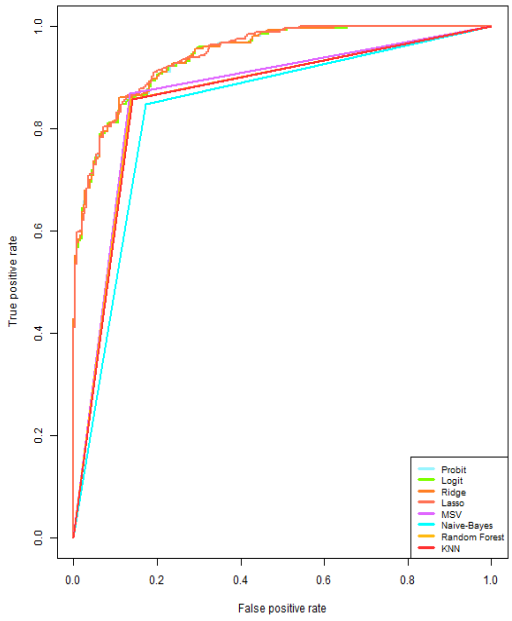
Debido al comportamiento anterior, el área bajo la curva de todas las regresiones es considerablemente mayor al área bajo la curva de los demás modelos. En primer lugar se encuentra la regresión Lasso y Ridge con un área de 0.944, seguido de la regresión Probit y Logit con un área de 0.943. En último lugar se muestra al modelo de Naive-Bayes con un área bajo la curva de 0.837.

En cuanto al cutoff, que es el umbral de decisión de clasificación, se puede ver que los modelos supervisados distintos a las regresiones presentan un cutoff de 1, ya que en este caso no hay un cambio gradual probabilístico, como si se presenta en las regresiones cuyo cutoff se encuentra entre 0.51 y 0.55

El comporamiento del cutoff se muestra de manera clara en el gráfico generado, donde los modelos supervisados distintos a las regresiones, muestran un movimiento rígido, es decir, presentan un punto de inflexión claramente donde pasa de una clasificación a la otra. Mientras que en las gráficas generadas por las regresiones este cambio es más suave.

Modelo	Area	Cutoff
Lasso	0.9449367	0.5150516
Ridge	0.9440006	0.5469762
Probit	0.9432744	0.5576502
Logit	0.9432744	0.5545375
MSV	0.8674897	1.0000000
Random Forest	0.8593561	1.0000000
KNN	0.8573953	1.0000000
Naive-Bayes	0.8375938	1.0000000

(a) *Área Curva Roc y Cutoff*



(b) *Curvas Roc*

Figura 5.19: *Conjunto de las Curvas Roc generada por los diferentes modelos*

5.6.2.5. Errores de predicción de las series de tiempo INMEX

El quinto elemento a revisar serán unas gráficas que representan la serie del tiempo del índice y los diferentes momentos en los cuáles los modelos tuvieron un error al predecir el movimiento del precio del índice. Dependiendo del modelo aumentará o disminuirá el número de puntos azules (las diferencias de subida o bajada) que aparecerán sobre la serie de tiempo en color naranja.

Estas gráficas se encuentran dibujadas en un fondo separado de manera mensual para que sea más fácil la ubicación de los momentos en los cuáles hubo discordancia entre el movimiento real y el movimiento pronosticado.

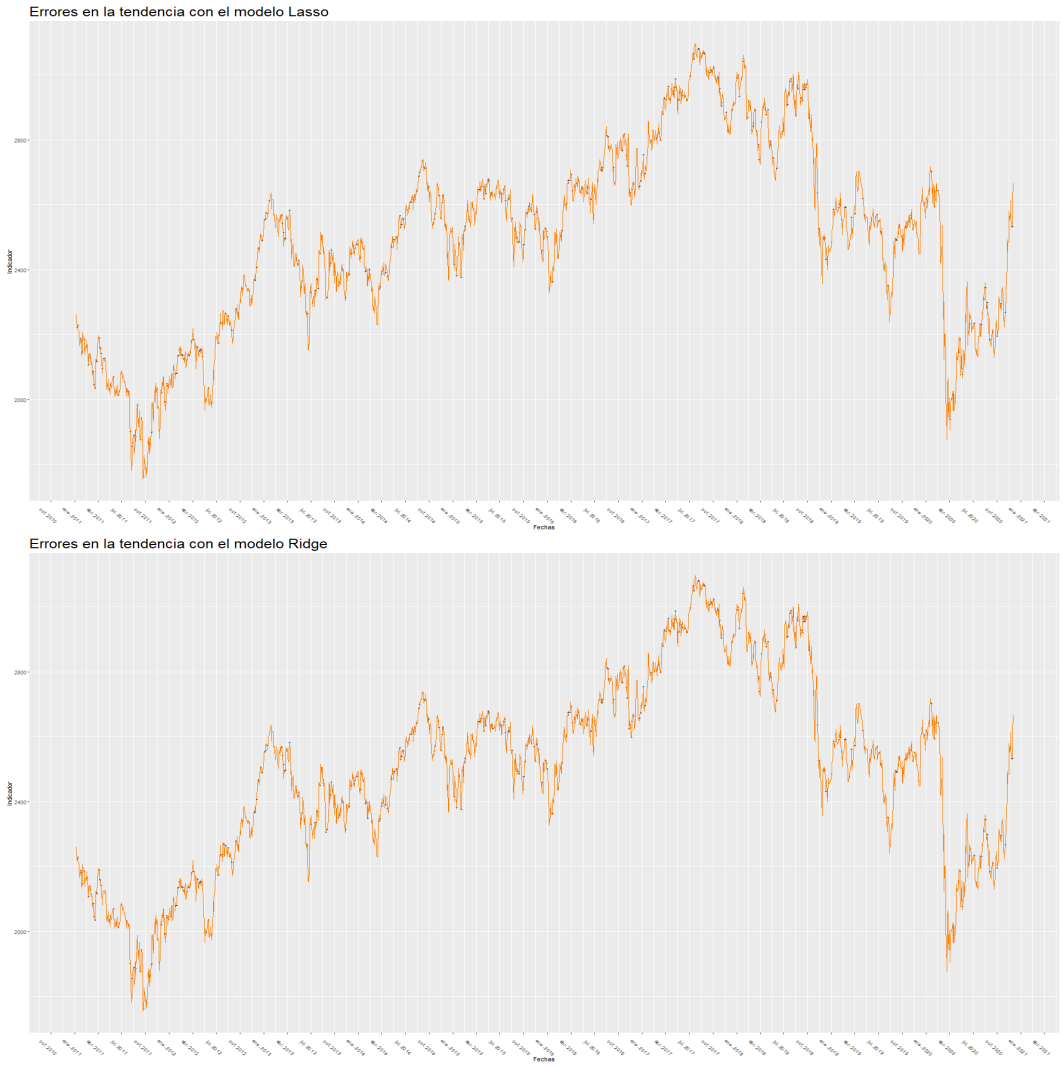


Figura 5.20: *Series de tiempo con diferencias en la predicción por parte de los modelos*

En la gráfica de la regresión Lasso, las diferencias entre las predicciones y los movimientos reales ocurren tanto en momentos de tendencia marcada, como en momentos con cambios de tendencia, sin embargo, se puede observar como es más complicado predecir los movimientos en momentos con mayor número de micromovimientos del mercado. Por otro lado, la regresión Ridge presenta problemas en la predicción cuando hay cambios en la tendencia, por pequeña que esta sea.

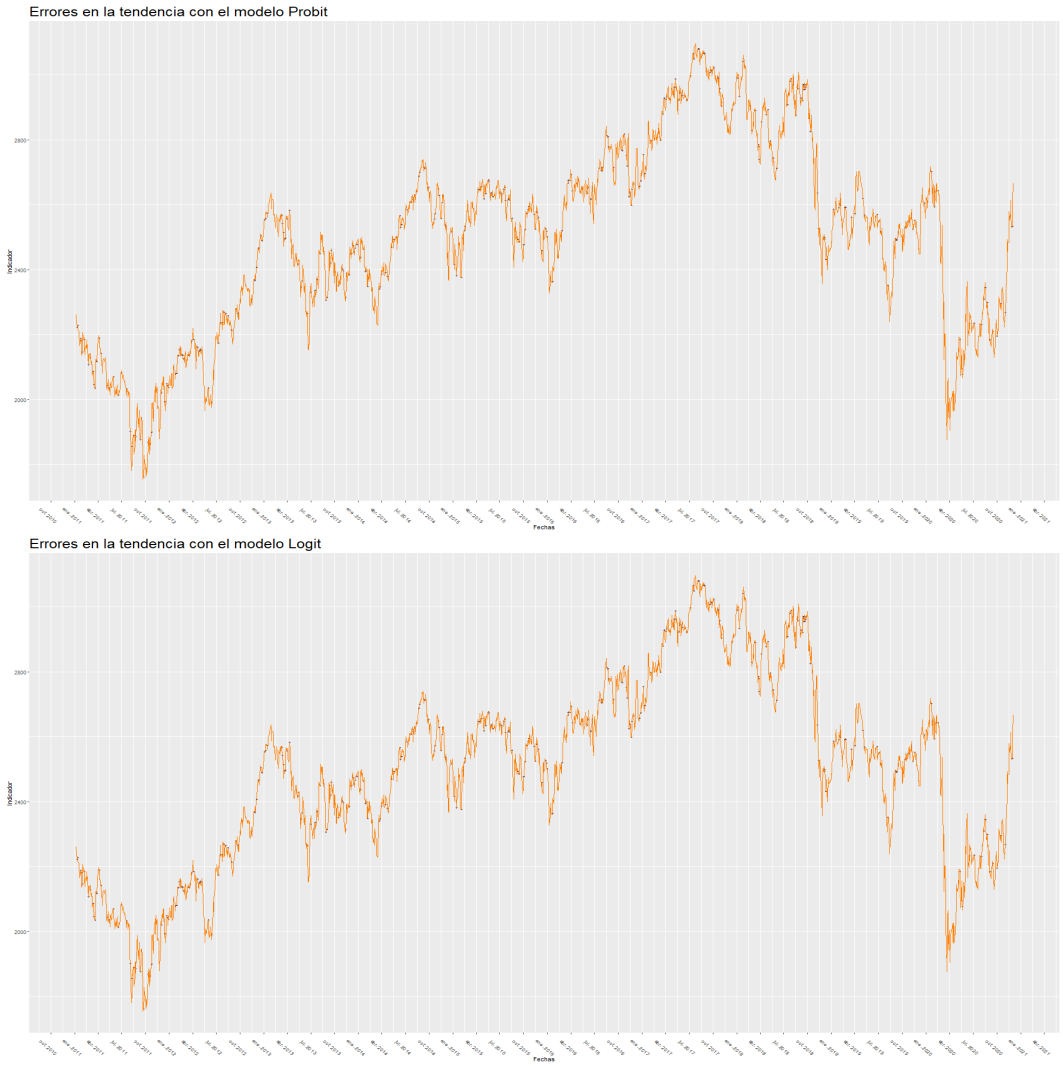


Figura 5.21: *Series de tiempo con diferencias en la predicción por parte de los modelos*

Haciendo una comparación entre la regresión Probit y Logit, se puede observar como en la ambas se presentan diferencias entre las predicciones y los movimientos reales, sin una tendencia visible, pues se encuentran errores tanto en las tendencias marcadas, como en los micromovimientos, siendo quizá un poco más visible en la gráfica Logit.

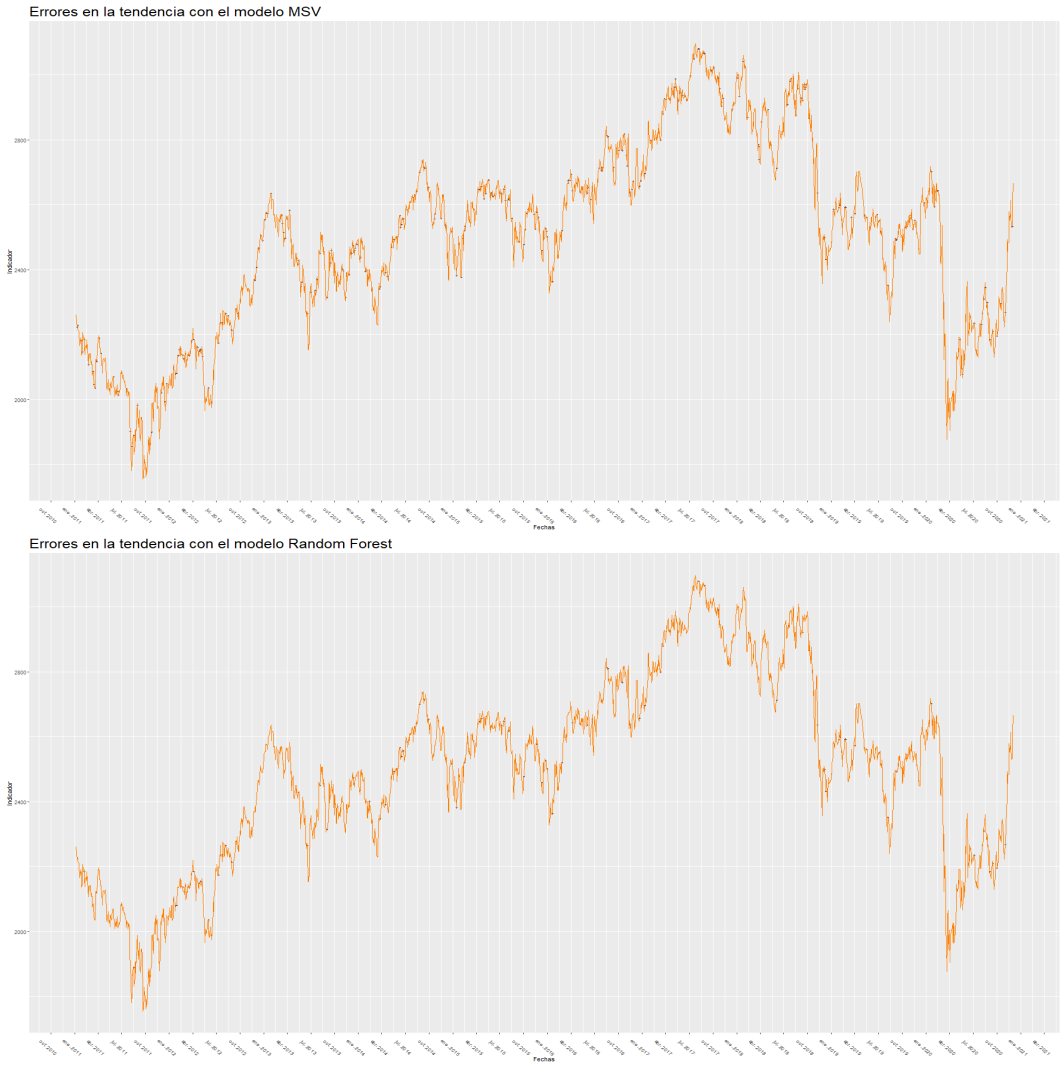


Figura 5.22: *Series de tiempo con diferencias en la predicción por parte de los modelos*

El comportamiento de las predicciones en las máquinas de soporte vectorial, se observa como la mayor cantidad de diferencias entre la predicción y los movimientos reales ocurre cuando se hace un cambio entre una tendencia alcista y una bajista, principalmente si existe una mayor volatilidad en el periodo. Por otra parte, la gráfica de random forest o bosque aleatorio presenta una menor cantidad de errores sobre esta gráfica, sin embargo, los errores ocurren cuando se tiene una tendencia sostenida y se llega a hacer una ligera corrección en los precios del índice.

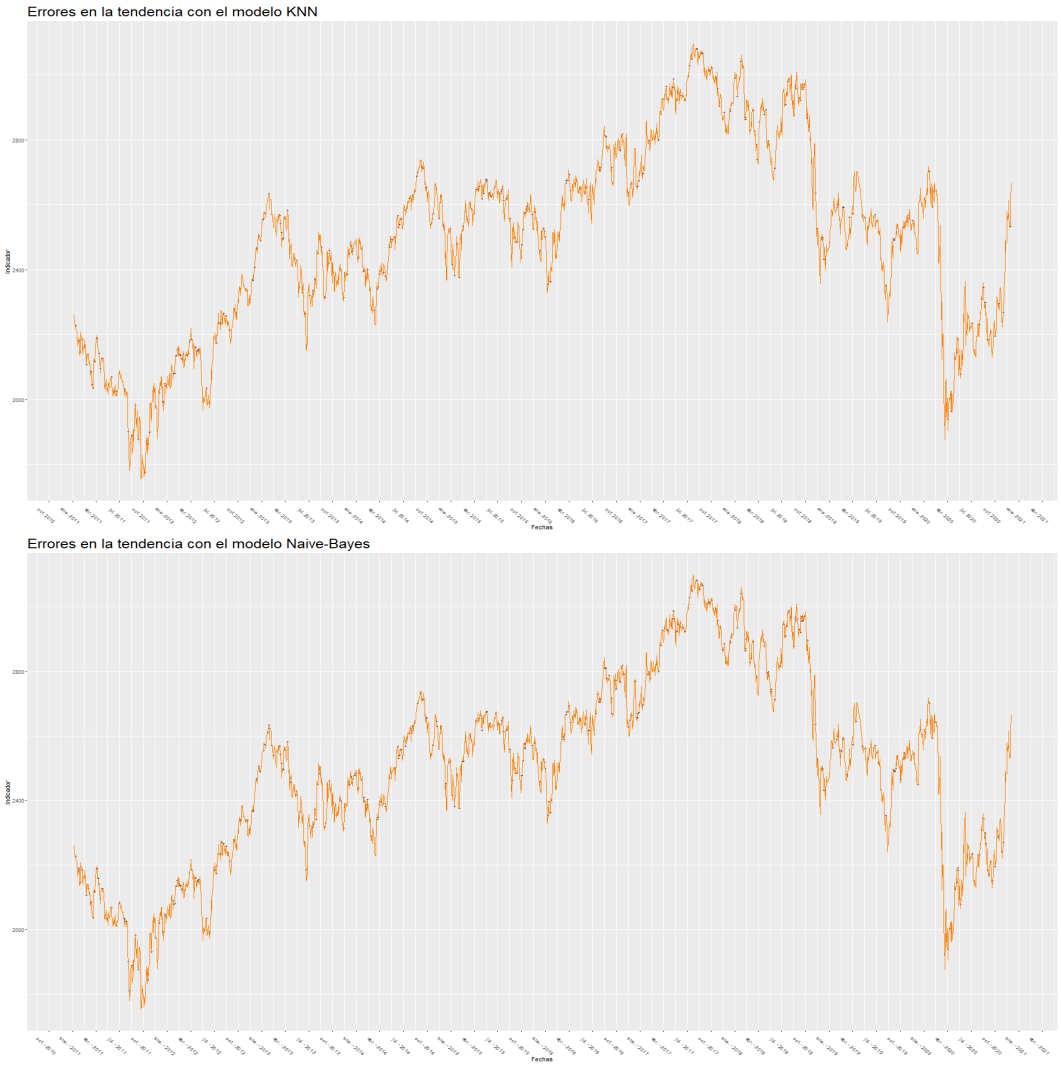


Figura 5.23: *Series de tiempo con diferencias en la predicción por parte de los modelos*

Las gráficas de KNN y Naive-Bayes, presentan una mayor cantidad de puntos azules sobre la serie de tiempo, indicando una mayor cantidad de dispariedades entre el movimiento predicho y el movimiento real en ese momento. En este caso, al igual que en las regresiones Logit y Probit no se encuentra como tal una tendencia en la distribución de los puntos sobre la serie de tiempo.

Cabe mencionar que en todas las mediciones numéricas, el modelo de Naive-Bayes presentó el peor comportamiento para predicción del movimiento del Inmex en el mercado.

5.6.2.6. Pruebas no paramétricas entre los diversos modelos para INMEX

El último conjunto de elementos para poder definir si hay un modelo mejor que otro para la predicción de movimientos en el Índice México son las tablas de los resultados de las diversas pruebas no paramétricas realizadas.

En en la tabla correspondiente a la prueba de Signos se puede ver que todos los elementos están en rojo, sin embargo, esto se puede atribuir a que dicha prueba no esta preparada del todo para clasificar elementos binarios, es decir, cuyas clases únicamente posean valores dicotómicos.

Por otro lado, la prueba de Wilcoxon presenta todos sus valores en verde, es decir, bajo esta prueba no se encuentran diferencias significativas entre el desempeño de los diferentes modelos ejecutados para el índice Inmex.

Cabe mencionar que estas pruebas se realizaron con un nivel de significancia del 0.95 %

Prueba Signos	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0	0	0	0	0	0	0
Naive-Bayes	0	0	0	0	0	0	NA
RandForest	0	0	0	0	0	NA	NA
MSV	0	0	0	0	NA	NA	NA
Ridge	0	0	0	NA	NA	NA	NA
Lasso	0	0	NA	NA	NA	NA	NA
Probit	0	NA	NA	NA	NA	NA	NA

Prueba Wilcoxon	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.8993	0.9496	1.0000	1.0000	0.9496	0.9496	0.704
Naive-Bayes	0.8001	0.7516	0.7040	0.7040	0.7516	0.6576	NA
RandForest	0.8494	0.8993	0.9496	0.9496	0.8993	NA	NA
MSV	0.9496	1.0000	0.9496	0.9496	NA	NA	NA
Ridge	0.8993	0.9496	1.0000	NA	NA	NA	NA
Lasso	0.8993	0.9496	NA	NA	NA	NA	NA
Probit	0.9496	NA	NA	NA	NA	NA	NA

Figura 5.24: *Prueba no paramétrica de Signos y Wilcoxon aplicado a los diferentes modelos de predicción*

Prueba Friedman	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.6698	0.8273	1.0000	1.0000	0.8185	0.8759	0.3304
Naive-Bayes	0.5164	0.4111	0.2888	0.3173	0.4658	0.3778	NA
RandForest	0.6473	0.7576	0.8815	0.8759	0.7237	NA	NA
MSV	0.8348	1.0000	0.8474	0.8348	NA	NA	NA
Ridge	0.1573	0.3173	1.0000	NA	NA	NA	NA
Lasso	0.5637	0.7630	NA	NA	NA	NA	NA
Probit	0.3173	NA	NA	NA	NA	NA	NA

Prueba Davenport	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	H0	H0	H0	H0	H0	H0	H0
Naive-Bayes	H0	H0	H0	H0	H0	H0	NA
RandForest	H0	H0	H0	H0	H0	NA	NA
MSV	H0	H0	H0	H0	NA	NA	NA
Ridge	H0	H0	H0	NA	NA	NA	NA
Lasso	H0	H0	NA	NA	NA	NA	NA
Probit	H0	NA	NA	NA	NA	NA	NA

Figura 5.25: *Prueba no paramétrica de Friedman y Davenport aplicado a los diferentes modelos de predicción*

Bajo las pruebas no paramétricas Friedman e Iman Davenport, se puede observar como todos los valores se muestran en verde, es decir, no presentan diferencias significativas entre los diversos modelos.

Rangos Alineados	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.8793	0.9394	1.0000	1.0000	0.9393	0.9407	0.6542
Naive-Bayes	0.7652	0.7087	0.6522	0.6535	0.7116	0.6108	NA
RandForest	0.8237	0.8818	0.9409	0.9407	0.8806	NA	NA
MSV	0.9396	1.0000	0.9398	0.9396	NA	NA	NA
Ridge	0.8769	0.9382	1.0000	NA	NA	NA	NA
Lasso	0.8781	0.9388	NA	NA	NA	NA	NA
Probit	0.9382	NA	NA	NA	NA	NA	NA

Prueba Quade	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.6703	0.8275	1.0000	1.0000	0.8188	0.8761	0.3309
Naive-Bayes	0.5170	0.4116	0.2893	0.3178	0.4664	0.3784	NA
RandForest	0.6478	0.7580	0.8817	0.8761	0.7241	NA	NA
MSV	0.8351	1.0000	0.8476	0.8351	NA	NA	NA
Ridge	0.1575	0.3178	1.0000	NA	NA	NA	NA
Lasso	0.5642	0.7634	NA	NA	NA	NA	NA
Probit	0.3178	NA	NA	NA	NA	NA	NA

Figura 5.26: *Prueba no paramétrica de Rangos Alineados y Quade aplicado a los diferentes modelos de predicción*

Bajo las pruebas no paramétricas Quade y Rangos Alineados, al igual que las pruebas anteriores, no muestran alguna diferencia significativa entre los modelos comparados, por lo que hasta este punto, se puede decir que para el índice Inmex, no hay un mejor modelo que otro.

Prueba	P_value
Friedman	0.9344
RangosAlin	0.9982
Quade	0.9346
Davenport	H0

Figura 5.27: *Pruebas no paramétricas múltiples aplicadas a los diferentes modelos de predicción*

Por último, se puede ver que las pruebas no paramétricas múltiples no detectaron ninguna diferencia significativa entre los resultados de los distintos modelos ocupados para la predicción de movimientos del Inmex. Por lo cuál, si bien puede existir alguna diferencia entre algunos elementos de manera individual, en el comportamiento conjunto de los resultados realmente no existe un modelo sobresaliente a los demás.

En este caso a diferencia del IPC, se puede concluir, a través de las pruebas no paramétricas, que no existen diferencias significativas entre los 8 modelos utilizados.

5.6.3. Resultados del Standard & Poors 500

El último índice que se ha de revisar será el S&P 500. Este índice de origen presenta un comportamiento muy diferente a los dos índices analisados anteriormente, pues desde su composición, diversificación y país de origen ya presenta diferencias.

El primer punto a revisar de los resultados obtenidos es el gráfico de pares, el cuál tiene la misma composición que el gráfico de IPC visto previamente, es decir, contiene una distribución de subidas y abajadas en la diagonal, un nivel de correlación en el triángulo superior y una nube de puntos en el triángulo inferior.

En el siguiente gráfico podemos observar:

1. Éste muestra un comportamiento más leptocúrtico en la distribución del momentum, así como en la distribución del índicador MACD. Por otro lado, la distribución de

la media móvil simple es más parecida a la vista por el IPC aunque con curvas más pronunciadas.

2. En las nubes de puntos que se presentan aquí, se puede ver que la concentración de puntos es ligeramente mayor que en los índices anteriores, es decir, las nubes son más compactas. Esto se puede apreciar más fácilmente en la relación entre los indicadores de Momentum y MACD con los demás indicadores.

3. Al igual que en las gráficas de pares anteriores, el Oscilador de Acumulación/Distribución sigue presentando una separación de los movimientos al alza y a la baja bastante claro en dos nubes de puntos distintas.

4. Otra similitud que presenta este gráfico con los vistos anteriormente, es la correlación que muestran los indicadores de RSI con el Momentum, el Estimador Estocástico K y con el Estimador Estocástico D.

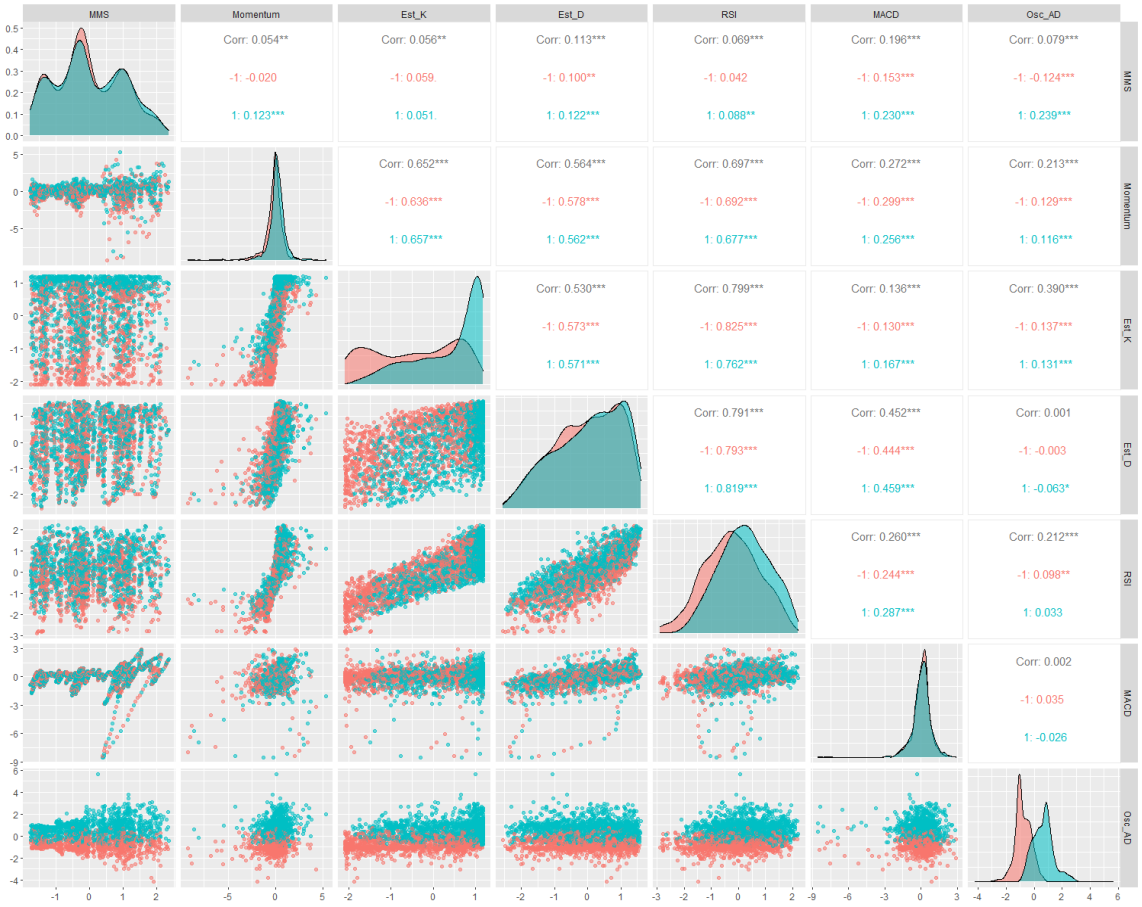


Figura 5.28: *GGPairs de los índicadores aplicados sobre el INMEX*

5.6.3.1. Variables de decisión del bosque aleatorio S&P 500

En esta sección se analizará brevemente cuáles fueron las variables que más tuvieron peso en la toma de decisión del bosque aleatorio para el índice IPC. Para ello se tomará en cuenta el resultado mostrado en las siguientes dos gráficas, donde en la primera de lado izquierdo se puede observar qué tanto decaería la predicción del modelo de Bosques Aleatorios si removiecemos alguno de los indicadores, mientras que en el gráfico de lado derecho se muestra la pureza de los nodos al final de los árboles de decisión si se removiece alguno de estos indicadores.

En la gráfica de "MeanDecreaseAccuracy" nuevamente se puede ver al Oscilador de Acumulación/Distribución como el indicador de mayor preponderancia en la decisión del Bosque Aleatorio. Por otra parte cabe resaltar en este punto, que el indicador con una

menor importancia en la precisión del Bosque Aleatorio es la Media Móvil Simple, pues al igual que como lo indica la gráfica de GGpairs que se vió anteriormente, no es muy bueno a la hora de diferenciar entre movimientos al alza o a la baja.

En término de pureza de nodo, como se ve en la gráfica de "MeanDecreaseGini", el Oscilador A/D y el Oscilador Estocástico K son los dos más importantes, nuevamente la Media Móvil Simple se encuentra en último lugar de importancia.

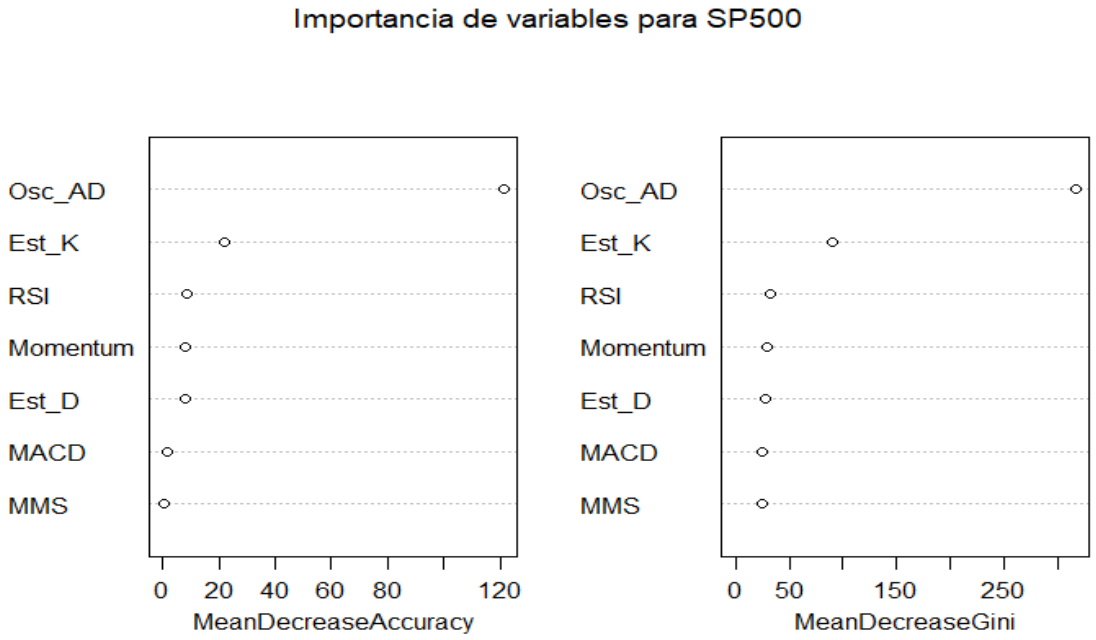


Figura 5.29: *Gráfica con las variables de mayor importancia para el Bosque Aleatorio del SP 500*

Por último, la siguiente tabla nos muestra una relevancia mayor en el uso del Oscilador A/D y del Oscilador Estocástico K como predictores para el Bosque Aleatorio del S&P 500, teniendo en cuenta que la diferencia de los dos primeros es bastante importante en relación a los demás indicadores.

Uso de variables por Bosque Aleatorio		
Indicador	MediaCaidaGini	Predictores usados
Oscilador AD	373.9467	19668
Estocástico K	87.3222	17069
RSI	30.1857	12477
Momentum	28.8843	11765
Estocástico D	25.5554	11681
MMS	25.5277	12137
MACD	25.3441	11675

5.6.3.2. Matriz de confusión S&P 500

En este caso, el comportamiento de los modelos llega a ser similar al comportamiento del índice IPC, dónde los primeros dos lugares estan ocupados por el modelo de bosque aleatorio y las máquinas de soporte vectorial. En el caso concreto del SP500, se tiene que bosques aleatorios tuvo un error de predicción de 57 elementos, mientras que las msv, tuvieron un error de predicción de 62 elementos.

Posteriormente se presentan las regresiones Ridge y Lasso, con un 66 elementos mal clasificados en ambos modelos. Por otra parte, los modelos que tuvieron un peor comportamiento fueron KNN y Naive-Bayes, dado que sus predicciones no fueron tan acertadas con 68 y 82 elementos mal clasificados. Es notable como Naive-Bayes tiene un desempeño más pobre que el penúltimo lugar.

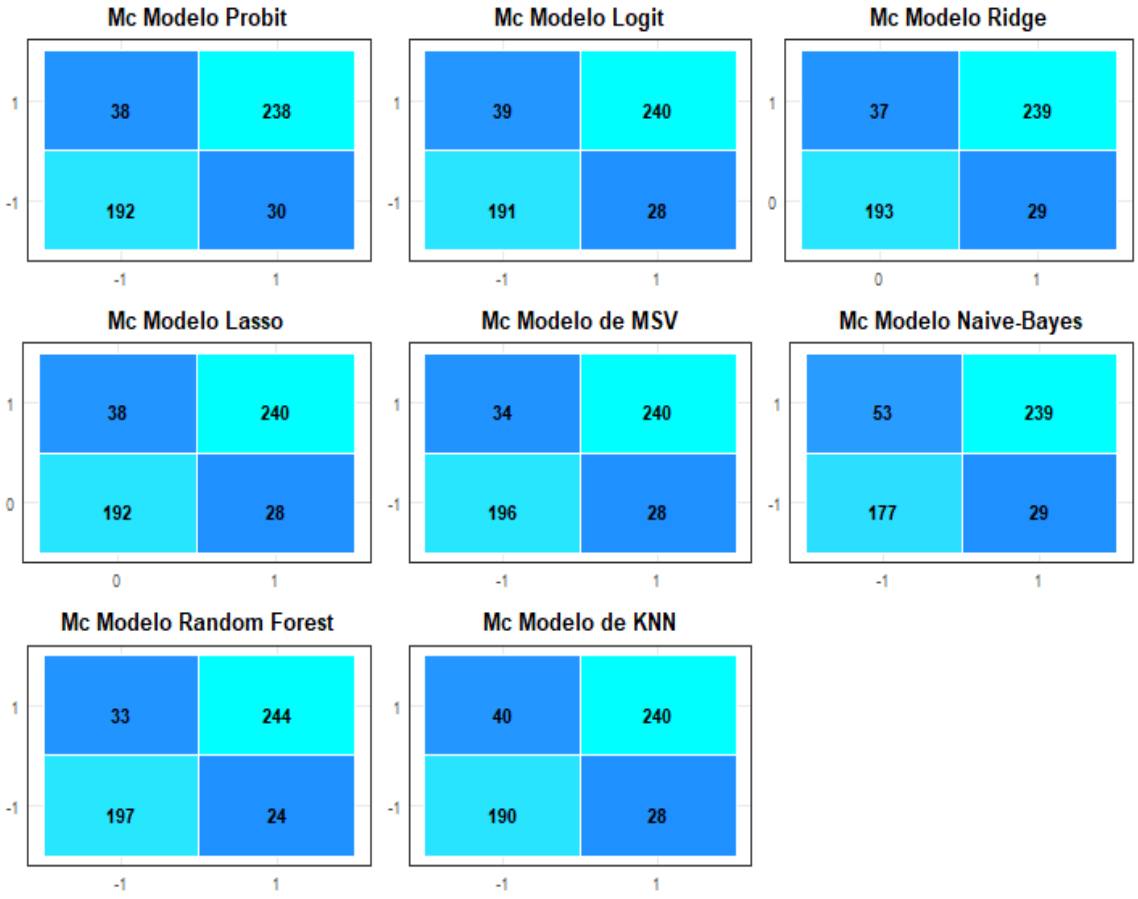


Figura 5.30: *Conjunto de las Matrices de Confusión de los modelos aplicados al IPC*

5.6.3.3. Métricas de clasificadores discretos S&P 500

Los resultados para el S&P 500 , se puede observar que el desempeño del modelo de bosques aleatorios es superior en todas las medidas que sus contrapartes. A su vez, el modelo de máquinas de soporte vectorial, resulta ser superior al resto de los modelos, exceptuando obviamente a bosques aleatorios.

Cabe resaltar que las medidas generadas por las regresiones Ridge, Lasso, Probit y Logit, llegan a ser muy similares, tanto en accuracy, sensibilidad, recall y F1. Aunque llegan a presentar una ligera diferencia en cuanto a su especificidad y precisión.

Por último, es claro que las medidas generadas por el modelo de Naive-Bayes llega a ser bastante inferior en relación a las regresiones y aún más cuando se compara con las msv o

el modelo de bosques aleatorios.

Modelo	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
Random Forest	0.8855422	0.8808664	0.8914027	0.9104478	0.8808664	0.8954128
MSV	0.8755020	0.8759124	0.8750000	0.8955224	0.8759124	0.8856089
Ridge	0.8674699	0.8659420	0.8693694	0.8917910	0.8659420	0.8786765
Lasso	0.8674699	0.8633094	0.8727273	0.8955224	0.8633094	0.8791209
Logit	0.8654618	0.8602151	0.8721461	0.8955224	0.8602151	0.8775137
Probit	0.8634538	0.8623188	0.8648649	0.8880597	0.8623188	0.8750000
KNN	0.8634538	0.8571429	0.8715596	0.8955224	0.8571429	0.8759124
Naive-Bayes	0.8353414	0.8184932	0.8592233	0.8917910	0.8184932	0.8535714

Figura 5.31: *Métricas derivadas de las matrices de confusión para el INMEX*

5.6.3.4. Curvas ROC S&P 500

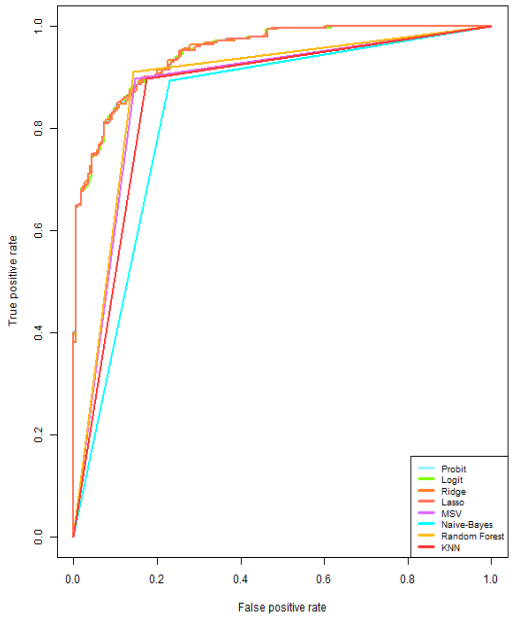
El cuarto elemento a analizar es el resultado de las curvas ROC. A continuación se muestrán dichos resultados tanto de manera gráfica como en forma de tabla, donde se podrá revisar el área bajo las distintas curvas, así como el cutoff de cada uno de los modelos.

El resultado generado por las curvas ROC llega a contrastar con los resultados vistos hasta ahora, pues muestra que las regresiones presentan una mayor área bajo la curva que los modelos supervisados. Esto puede deberse a la propia forma de las curvas, ya que las regresiones por su construcción probabilística van generando una gráfica más suave, abarcando un mayor espacio. Por otra parte, al igual que el comportamiento visto con los otros dos índices, los modelos supervisados al tener un cutoff de 1, cambian de dirección de manera única formando una especie de cuadrado, y haciendo que el área del mismo sea menor.

Adicionalmente, y como ya se vió en las gráficas, el cutoff de las curvas llega a ser menor a uno, lo que propicia que el diseño de éstas sea más sueave que aquellos modelos que presentan un cutoff de 1, haciendolas "puntiagudas".

Modelo	Area	Cutoff
Ridge	0.9491239	0.4930998
Lasso	0.9490266	0.6157404
Probit	0.9489942	0.6184278
Logit	0.9489293	0.6193964
Random Forest	0.8834848	1.0000000
MSV	0.8738482	1.0000000
KNN	0.8608047	1.0000000
Naive-Bayes	0.8306781	1.0000000

(a) *Área Curva Roc y Cutoff*



(b) *Curvas Roc*

Figura 5.32: *Conjunto de las Curvas Roc generada por los diferentes modelos*

5.6.3.5. Errores de predicción de las series de tiempo S&P 500

Las gráficas que representan las series de tiempo del índice y los puntos que representan los errores de predicción con los datos completos, serán el quinto elemento a analizar para el S&P 500. Al igual que para los otros dos índices, la cantidad de puntos azules sobre la gráfica podrían aumentar o disminuir respecto al modelo utilizado.

Estas gráficas se encuentran dibujadas en un fondo separado de manera mensual para que sea más fácil la ubicación de los momentos en los cuáles hubo discordancia entre el movimiento real y el movimiento pronosticado.

Adicionalmente, cabe mencionar que el propio comportamiento de la serie del tiempo del índice es diferente a los índices anteriores, ya que se trata de una conformación y regionalidad distinta a los anteriores dos índices.

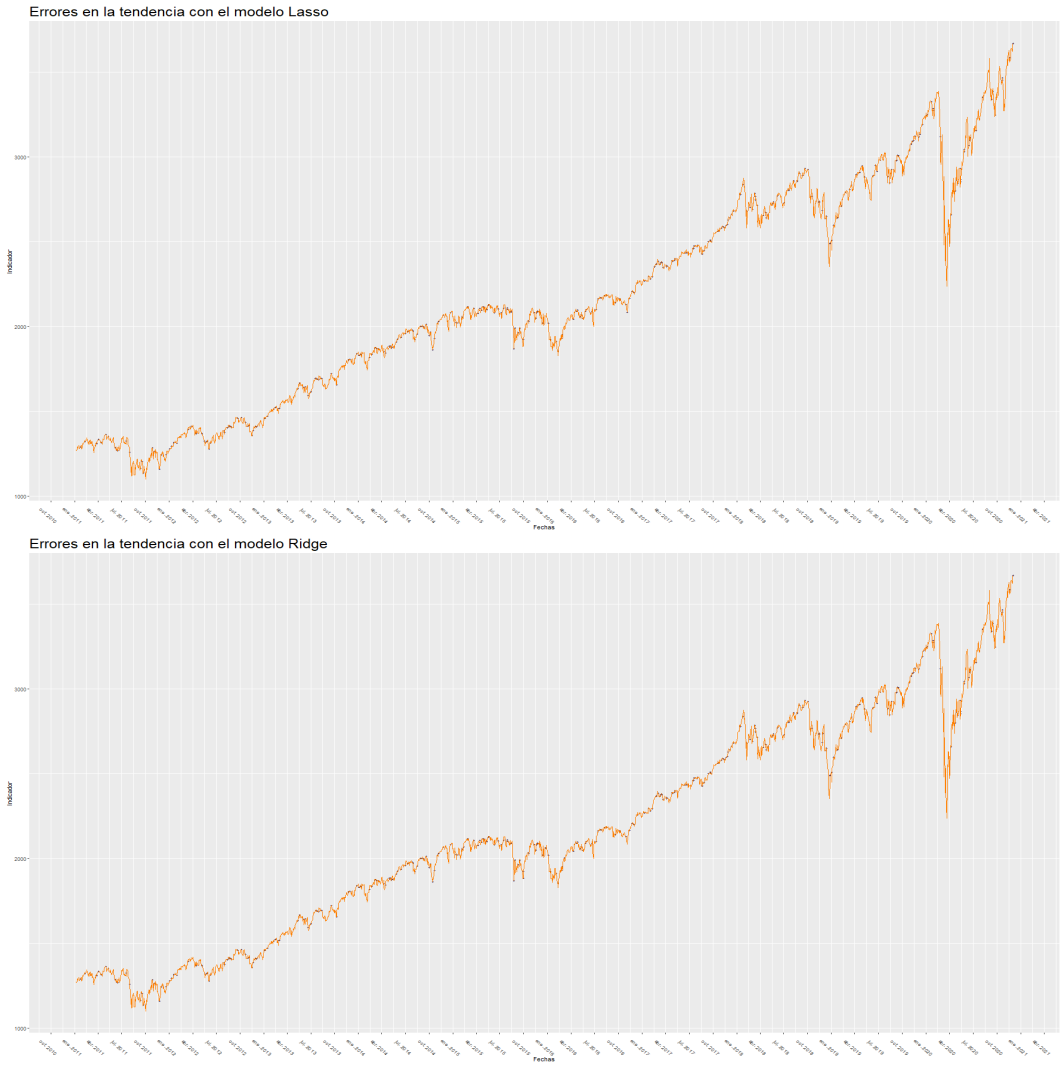


Figura 5.33: *Series de tiempo con diferencias en la predicción por parte de los modelos*

En el caso de las diferencias entre la predicción y el movimiento real para la gráfica Ridge, se puede observar como "le cuesta más trabajo", o es menos precisa la predicción del modelo Ridge, en momentos de alta volatilidad en periodos cortos de tiempo, es decir, en los micromovimientos del mercado. Sin embargo, en tendencias marcadas como lo son la caida en febrero del 2020 o bien la posterior recuperación, se presentan un número menor de errores.

Por otra parte la regresión Lasso no sigue este mismo comportamiento, al presentar una mayor cantidad de errores tanto en los micromovimientos como en las tendencias marcadas.

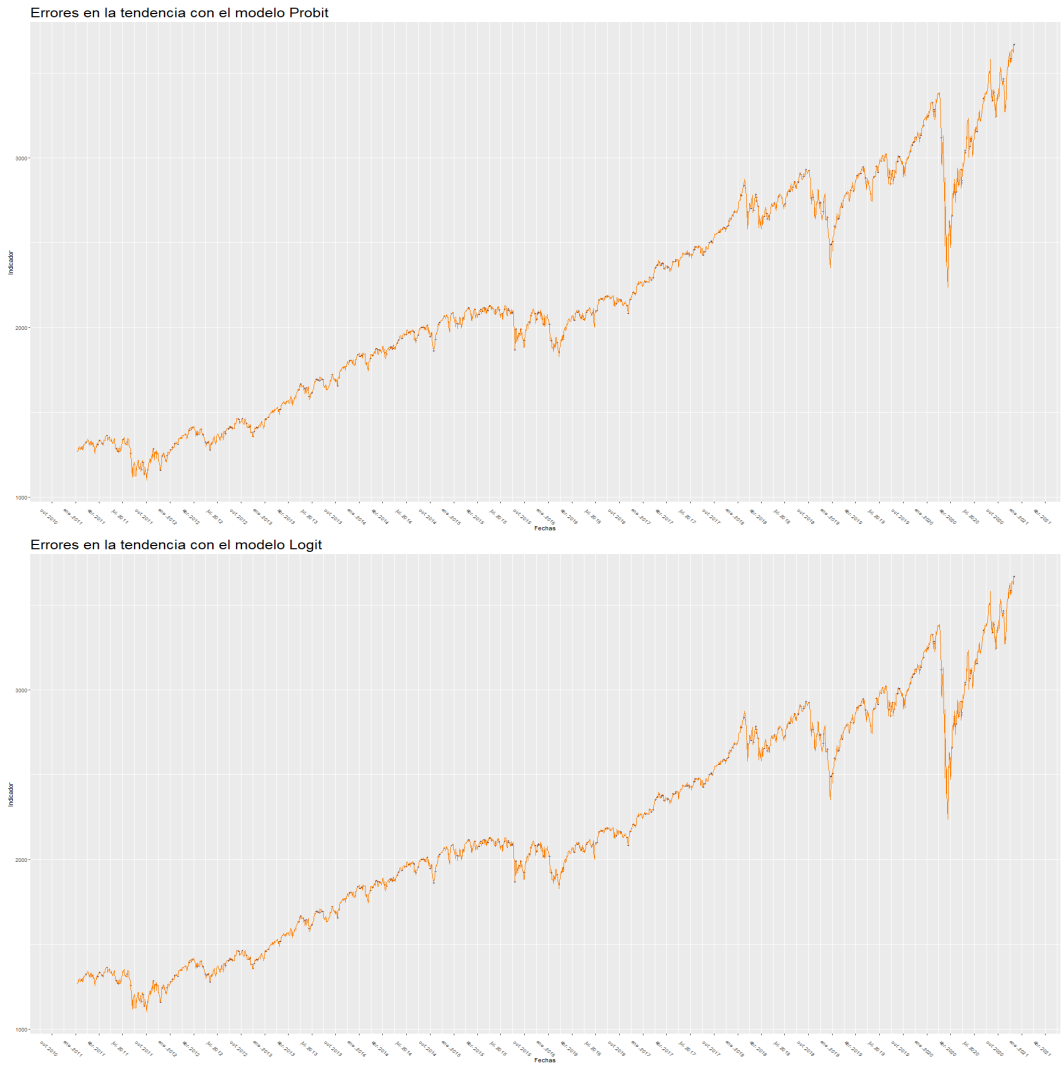


Figura 5.34: *Series de tiempo con diferencias en la predicción por parte de los modelos*

El comportamiento de las gráficas de diferencias entre la regresión Logit y Probit son muy similares, inclusive con la regresión Lasso, donde se presentan algunas diferencias tanto en micromovimientos, como en tendencias marcadas, como es visible en el periodo de febrero del 2014 a octubre del 2016, y en la caida del mercado en febrero del 2020 y su posterior recuperación.

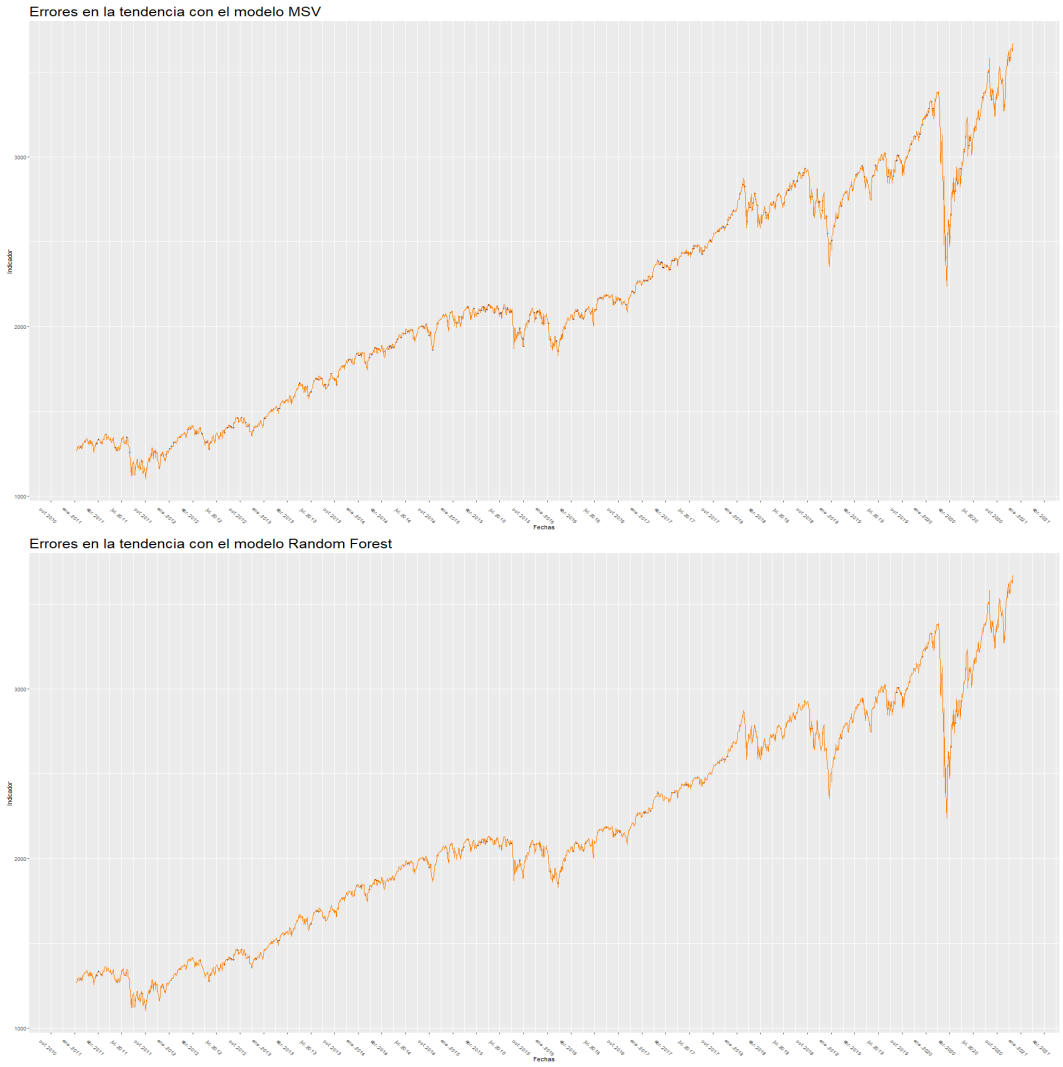


Figura 5.35: *Series de tiempo con diferencias en la predicción por parte de los modelos*

En cuanto a la relación entre el modelo de MSV y Random Forest, se puede apreciar con facilidad que el rendimiento del segundo es mucho mejor con los datos completos que las MSV. En este sentido no se podría hablar de que Random Forest presente una sensibilidad mayor por micromovimientos, como en los otros dos índices, ya que en general el S&P 500 presenta una tendencia de movimientos de subida, al menos con una mayor intensidad que aquellos de bajada, como se pudo observar en la tabla de S&P de la sección 4 de este capítulo.

En cuanto a las MSV, se puede observar que la mayoría de los errores se presentan en los micromovimientos de pequeñas tendencias, principalmente en los picos de éstas.

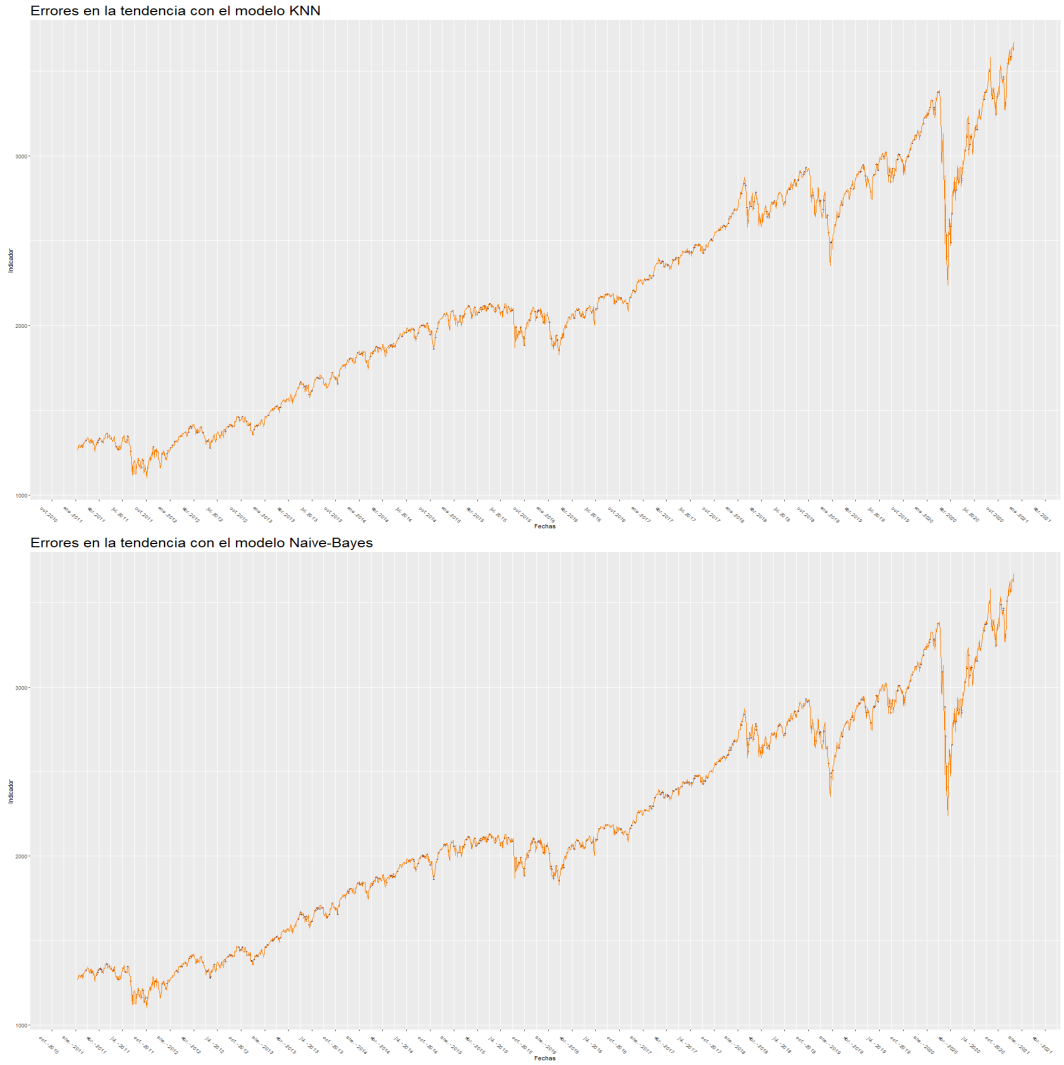


Figura 5.36: *Series de tiempo con diferencias en la predicción por parte de los modelos*

Naive-Bayes se puede ver ligeramente más densa que la gráfica de KNN, además presenta una concentración uniforme de errores entre las predicciones y los movimientos reales, pues no parece haber una tendencia marcada ya sea en micromovimientos o bien en movimientos de mayor duración.

Por su parte, KNN presenta una mayor facilidad en predecir movimientos cuando se encuentran en medio de una tendencia ya sea al alza o a la baja, y presenta una mayor dificultad en momentos de mayor volatilidad.

5.6.3.6. Pruebas no paramétricas entre los diveros modelos para S&P 500

El sexto elemento a considerar en el análisis sobre si existe un mejor modelo en comparación a otro para la predicción de movimientos en el Standard & Poors 500, son los resultados de las pruebas no paramétricas aplicadas a los diferentes modelos. Estas pruebas fueron hechas con una confianza del 95 %.

La primer tabla es la prueba de signos donde se puede ver que todos los elementos están en rojo, empero, esto se puede deber a que dicha prueba no considera en su cálculo a variables dicotómicas y esto entorpece el cálculo de la prueba, y por ende la fuerza de la misma.

La siguiente prueba que se observa, es la de Wilcoxon, donde se puede observar que ningún elemento se encuentra marcado en rojo, lo que implica que ningún elemento presenta alguna diferencia significativa entre los modelos pares.

Prueba Signos	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0	0	0	0	0	0	0
Naive-Bayes	0	0	0	0	0	0	NA
RandForest	0	0	0	0	0	NA	NA
MSV	0	0	0	0	NA	NA	NA
Ridge	0	0	0	NA	NA	NA	NA
Lasso	0	0	NA	NA	NA	NA	NA
Probit	0	NA	NA	NA	NA	NA	NA

Prueba Wilcoxon	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.9480	1.0000	0.7441	0.7441	0.8961	0.6015	0.2377
Naive-Bayes	0.2647	0.2377	0.1318	0.1318	0.1898	0.0887	NA
RandForest	0.5568	0.6015	0.8449	0.8449	0.6954	NA	NA
MSV	0.8447	0.8961	0.8448	0.8448	NA	NA	NA
Ridge	0.6952	0.7441	1.0000	NA	NA	NA	NA
Lasso	0.6952	0.7441	NA	NA	NA	NA	NA
Probit	0.9480	NA	NA	NA	NA	NA	NA

Figura 5.37: *Prueba no paramétrica de Signos y Wilcoxon aplicado a los diferentes modelos de predicción*

Prueba Friedman	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.8348	0.3938	0.6698	0.3711	0.2733	0.5900	0.0516
Naive-Bayes	0.0526	0.0183	0.0390	0.0136	0.0181	0.0469	NA
RandForest	0.7237	0.8575	0.8575	0.8527	0.5485	NA	NA
MSV	0.3173	0.6831	0.4142	0.6698	NA	NA	NA
Ridge	0.0833	1.0000	0.3173	NA	NA	NA	NA
Lasso	0.3173	0.1573	NA	NA	NA	NA	NA
Probit	0.0833	NA	NA	NA	NA	NA	NA

Prueba Davenport	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	H0	H0	H0	H0	H0	H0	H0
Naive-Bayes	H0	H1	H1	H1	H1	H1	NA
RandForest	H0	H0	H0	H0	H0	NA	NA
MSV	H0	H0	H0	H0	NA	NA	NA
Ridge	H0	H0	H0	NA	NA	NA	NA
Lasso	H0	H0	NA	NA	NA	NA	NA
Probit	H0	NA	NA	NA	NA	NA	NA

Figura 5.38: *Prueba no paramétrica de Friedman y Davenport aplicado a los diferentes modelos de predicción*

El siguiente par de tablas a revisar son Friedman y Davenport, sin embargo, a diferencia de las tablas anteriores o incluso a diferencia de los índices ya vistos, se puede observar como Naive-Bayes presenta una diferencia significativa contra todos los demás modelos pares en ambas pruebas, exceptuando la regresión Logit. Esto significa que existe una consistencia no solo en el comportamiento de la prueba, sino también en el rendimiento de todos los modelos en comparación a Naive-Bayes.

Analizando esto de manera independiente no se sabría si este modelo es mejor o peor a

los demás, empero, con el resto de los elementos revisados hasta ahora se podría dar una mejor interpretación de este resultado.

Rangos Alineados	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.9396	0.7614	0.8793	0.7610	0.6515	0.8215	0.3703
Naive-Bayes	0.3353	0.2362	0.3000	0.2343	0.1883	0.2724	NA
RandForest	0.8806	0.9401	0.9401	0.9399	0.8204	NA	NA
MSV	0.7052	0.8796	0.7619	0.8793	NA	NA	NA
Ridge	0.8164	1.0000	0.8771	NA	NA	NA	NA
Lasso	0.9382	0.8769	NA	NA	NA	NA	NA
Probit	0.8164	NA	NA	NA	NA	NA	NA

Prueba Quade	Logit	Probit	Lasso	Ridge	MSV	RandForest	Naive-Bayes
KNN	0.8351	0.3943	0.6703	0.3716	0.2738	0.5905	0.0515
Naive-Bayes	0.0525	0.0182	0.0389	0.0134	0.0180	0.0468	NA
RandForest	0.7241	0.8577	0.8577	0.8529	0.5490	NA	NA
MSV	0.3178	0.6835	0.4148	0.6703	NA	NA	NA
Ridge	0.0833	1.0000	0.3178	NA	NA	NA	NA
Lasso	0.3178	0.1575	NA	NA	NA	NA	NA
Probit	0.0833	NA	NA	NA	NA	NA	NA

Figura 5.39: *Prueba no paramétrica de Rangos Alineados y Quade aplicado a los diferentes modelos de predicción*

En cuanto a los resultados arrojados por la prueba de Rangos Alienados, se puede ver que no se han detectado diferencias significativas entre el comportamiento de los modelos, sin embargo, los resultados de la prueba de Quade vuelven a encontrar una diferencia significativa en el rendimiento del modelo de Naive-Bayes y el resto de los modelos, salvo la regresión Logit.

En relación al resultado de Naive-Bayes es de llamar la atención que se vuelva a presentar el mismo resultado que en las pruebas de Friedman y Davenport.

Prueba	P_value
Friedman	0.0209
RangosAlin	0.6946
Quade	0.0208
Davenport	H1

Figura 5.40: *Pruebas no paramétricas múltiples aplicadas a los diferentes modelos de predicción*

Por último, se puede ver que en la prueba no paramétrica múltiple de Fridman y Quade presentan un p-value muy pequeño, dado que se tiene una prueba de significancia del 0.95, adicionalmente, la prueba de Iman Davenport también presenta un rechazo a la hipótesis nula, es decir, si se presenta una diferencia significativa entre el desempeño de los múltiples modelos. Sin embargo, más adelante se analizará si realmente se puede decir si existe un modelo mejor que los demás.

5.7. Elección del mejor modelo

Tomando en cuenta los resultados obtenidos por las matrices de confusión, se puede observar como las máquinas de soporte vectorial se encuentran presentes en los primeros lugares de los tres índices, con un accuracy de 0.8709 para el IPC, 0.8674 para el INMEX y un 0.8755 para el S&P 500. También otro modelo que se encuentra presente como primer lugar de accuracy en dos de los tres índices es el modelo de bosques aleatorios, con valores iguales a 0.8951 para el IPC y de 0.8855 para el S&P 500. Para el caso del INMEX, las MSV fueron el primer lugar, seguido de la regresión Lasso con 0.8654 de accuracy.

Sin embargo, como se mencionó al inicio de este capítulo, las matrices de confusión eran solo uno de los múltiples elementos a analizar para poder decidir cuál de estos modelos sería el mejor. Si se toma en cuenta el área de la curva ROC, se tiene que las regresiones Ridge y Lasso, también se desempeñan de una manera bastante adecuada, al encontrarse en el tercer lugar de accuracy para los modelos y además estar en primer lugar en las mediciones de la curva ROC, con un área superior a otros modelos.

Por último, hay que considerar que realmente las pruebas no paramétricas no señalaron que hubiera una diferencia significativa con algunos de los modelos mencionados en los párrafos anteriores, salvo el caso del modelo de Naive-Bayes cuyo desempeño ha sido el peor para todos los índices analizados, por lo que se podría decir ya en este punto, es que si bien existe una diferencia entre este modelo y los demás, no necesariamente implica que sea por su buena predicción con los datos analizados.
w3

Si bien es cierto que ningún modelo fue superior en todos los ámbitos para todos los índices, si se puede tener una idea de aquellos que sobresalieron de sus páres en algunas o varias de los análisis hechos, por lo que el siguiente elemento a considerar serían los pros y los contras de cada uno de los modelos revisados hasta ahora.

5.7.1. Regresión regularizada Ridge

Ventajas:

1. La principal ventaja que posee este clasificador es la reducción de la varianza. Sin embargo, esto lo hace a cambio de un mayor sesgo. Es decir, en presencia de colinearidad es mejor tener un resultado sesgado, para poder obtener una mejor varianza.

2. Otra ventaja que presenta este método, al igual que lo hacen el resto de regresiones, es que son fáciles de calcular y por ende su entrenamiento no resulta costoso computacionalmente hablando.

Desventaja:

1. Poseer un sesgo grande.

2. El modelo final incluirá a todos los predictores, pues si bien la penalización los acerca al cero, nunca llegarán a ser exactamente cero.

5.7.2. Regresión regularizada Lasso

Ventajas:

1. A diferencia de Ridge, una de las ventajas de esta regresión es que puede eliminar a los predictores con menor poder predictivo, de esta manera se obtienen los "mejores elementos" para predecir el fenómeno a estudiar.

Desventajas:

1. La principal desventaja de este método irónicamente es su principal ventaja, ya que en algunos casos la reducción de predictores puede ser tan grande, que elimine variables que pueden llegar a ser importante o interesante analizar.

2. La selección de parámetros puede tener un gran sesgo, ya que al utilizarlo en diferentes grupos de datos dentro de validación cruzada, sus características pueden llegar a ser bastante diferentes.

5.7.3. K-vecinos cercanos

Ventajas:

1. Cómo se mencionó en el capítulo 1, una de las grandes ventajas de este método es su simplicidad al no involucrar una gran cantidad de matemáticas más allá de la medición de las distancias. Por ende esto lo hace computacionalmente ligero de calcular.

2. Otra ventaja que presenta este método, es que permite hacer límites de decisión complicados debido a la generación de curvas suaves a partir del cálculo de funciones lineales por partes.

Desventajas:

1. Si una muestra no se encuentra correctamente equilibrada, es decir, existen relativamente los mismos números de elementos de ambas clases, entonces ocurrirá que los K vecinos más cercanos corresponderán a la clase con el tamaño más grande de muestra.

2. En volumenes grandes de información, este método puede llegar a ser computacionalmente costoso.

5.7.4. Máquinas de soporte vectorial

Ventajas:

1. Las MSV pueden ser bastante útiles cuando el conjunto de datos es no linealmente separable, de esta manera, se elige un núcleo que sea no lineal y se puede llevar a cabo la clasificación del conjunto de datos.

2. Otra ventaja que puede observarse en este método de clasificación es cuando se quiere clasificar elementos en dimensiones muy altas, pues el comportamiento de este modelo puede ser bastante bueno aún en altas dimensiones.

   En el caso específico de este trabajo, se puede observar como al menos en dos modelos se presentan como un método de clasificación bastante preciso al tener tasas de precisión arriba de 0.87.

Desventajas:

1. Los cálculos que se necesitan realizar el entrenamiento del modelo de MSV y obtener los mejores hiperparámetros para el conjunto de datos en cuestión, puede llegar a ser bastante tardado. Esto podría solucionarse quizá con un hardware más poderoso, sin embargo, no garantiza que se obtenga el resultado en un tiempo adecuado. De hecho, para el objetivo de este trabajo, donde se pretende pronosticar los movimientos del mercado de un día a otro, este método de clasificación no es funcional, al arrojar los resultados de entrenamiento en mínimo 2 días.

5.7.5. Bosque aleatorio o random forest

Ventajas:

1. Como se mencionó en el capítulo 1, en caso de que el conjunto de datos posea una gran no linalidad y una compleja relación entre las características y la respuesta, entonces el comportamiento del método de Bosques Aleatorios será bastante útil y bueno para dicho conjunto de datos.

2. Otra de sus ventajas es que este tipo de clasificadores se comportan bien en espacios de alta dimensión, así como con un gran número de elementos de entrenamiento.

Desventajas:

1. Al tener una gran varianza, este tipo de clasificador no es tan robusto como otros, ya que un pequeño cambio en la información de inicio puede causar un gran cambio en la estimación final del árbol.

5.7.6. Regresión Logística (Logit)

Ventajas:

1. Es un método de clasificación que puede ser ocupado como etapa inicial en el conjunto de datos que se tiene, al presentar en general un buen comportamiento de clasificación. Adicionalmente, computacionalmente no es un método costoso aún en grandes volumenes de datos.

Desventajas:

1. Es principal desventaja que presenta este método de clasificación es que la información a clasificar posea características de linealidad, pues de otra forma, el problema no se podrá resolver de manera directa con este método de clasifación.

5.7.7. Regresión Probit

Ventajas:

1. Al igual que la regresión logística, este método de clasificación es bastante sencillo y por ende, computacionalmente poco costoso aún en conjuntos grandes de datos. De igual manera, pese a que es un método "sencillo", el desempeño de este clasifidor generalmente presenta buenos resultados.

Desventajas:

1. Los coeficientes generados no presentan una interpretación directa.
2. De igual manera, requiere una cantidad razonablemente grande de información para hacer una estimación de máxima verosimilitud adecuada.

Cabe mencionar que en caso de elegir algún modelo ganador por sobre todos los índices tendrían que ser las regresiones regularizadas tanto Ridge como Lasso, en caso de dejar estas a un lado y tener que elegir otro clasificador, una buena opción sería la Máquina de Soporte Vectorial aunque poco práctica en la aplicación diaria.

Conclusiones

A través del desarrollo de este trabajo, se pudieron revisar algunas características de los índices accionarios del mercado financiero mexicano, tales como su composición y un poco de su historia. De igual manera, se pudo revisar como algunos hechos geo políticos y macroeconómicos llegaron a afectar su desempeño. En relación a esto, se observó cómo su comportamiento fue muy similar en el tiempo, pues ambos dependen de una misma economía. Sin embargo, a la hora de ejecutar las técnicas de clasificación, algunos de los resultados de éstos fueron un tanto diferentes.

De igual manera se pudo observar cómo fue el comportamiento de uno de los principales índices de economías desarrolladas, si bien hay algunos sucesos que llegaron a afectar a los índices nacionales como extranjeros, se puede ver que el impacto de estos sucesos geopolíticos tuvo una magnitud diferente en ambas economías. Esto quedó más claro a la hora de ejecutar los modelos de independencia vistos durante el capítulo 2, con resultados que mostraban cierta correlación entre los tres índices.

De las diversas técnicas que se llegaron a utilizar en este trabajo, las mejores sin duda fueron las máquinas de soporte vectorial y los bosques aleatorios. Esto tomando en cuenta la métricas generadas por las matrices de confusión, en especial el accuracy. De igual manera, cabe resaltar que las regresiones Ridge, Lasso y Probit también tuvieron un buen desempeño a la hora de clasificar, y mejor aún, a diferencia de los modelos supervisados las regresiones presentaron una mejor curva ROC y una mayor área bajo la curva.

Por otra parte, el análisis gráfico de los errores de pronóstico contra los movimientos reales, revela que la distribución de errores de las regresiones no necesariamente esta enfocado en algún sitio en particular, es decir, se pueden encontrar en micromovimientos, tendencias marcadas, o momentos de alta volatilidad. En el caso de las máquinas de soporte vectorial es apreciable como tiene dificultad para hacer una predicción correcta en momentos de alta volatilidad, sin embargo, el modelo de bosques aleatorios tiene un mejor comportamiento en momentos de alta volatilidad y micromovimientos; pero tiene un menor poder predictivo cuando después de una tendencia sostenida, se revierte dicho comporta-

miento.

De lo anterior se puede mencionar que si bien el accuracy de las regresiones no es tan elevado como pueden ser otros modelos, esto lo compensa con el desempeño ante la curva ROC y más importante, ante el tiempo de ejecución realizado, pues es considerablemente menor al del resto de los modelos, en especial MSV y bosques aleatorios.

Por último, vale la pena mencionar que el comportamiento del modelo de Naive-Bayes si bien no se puede decir que sea malo en sí mismo, comparando sus resultados con los demás modelos, su poder de predicción llega a ser menor. Inclusive, se puede pensar que para algunos índices, llega a existir una diferencia significativa entre su desempeño con el resto de los modelos, como consecuencia de los resultados obtenidos a través de las diversas pruebas no paramétricas.

En contraste, el comportamiento del resto de las técnicas de aprendizaje máquina llegó a ser ligeramente diferente, incluso en los índices más parecidos. En algunos casos se observó cómo algunas técnicas, como lo son Bosques Aleatorios, fueron más precisas ante micromovimientos y menos precisas ante tendencias sostenidas. Adicionalmente, un resultado que al igual que las regresiones regularizadas fue consistente para todos los índices, fue el método de Naive-Bayes, el cual presentó los resultados menos precisos de clasificación.

Si bien la mayoría de las técnicas ejecutadas en la sección de entrenamiento a través del método de validación Cruzada fueron entrenadas rápidamente, hubo dos en particular cuyo desempeño fue bastante lento. Estas técnicas son el bosque aleatorio y las máquinas de soporte vectorial. Las MSV obtuvieron un resultado considerablemente bueno para los tres índices estudiados, sin embargo, el tiempo de entrenamiento que tuvieron las vuelve poco prácticas al tardar entre dos y tres días en obtener los resultados. Es probable que haciendo una reducción en los hiperparámetros a estimar por VC y con un hardware un poco más poderoso, estos tiempos puedan reducirse considerablemente.

Esta observación, también da pie a mencionar que si bien se consideró utilizar alguno de los diversos métodos de validación cruzada múltiple con la finalidad de generar no solamente una predicción más robusta, sino también poder dar mayor sustento a la selección del mejor modelo, esta idea fue descartada para los fines de este trabajo, ya que no se cuenta con la infraestructura necesaria en la nube para poder montar dicho proceso; y que de llega a ser montado en una computadora convencional de escritorio o portatil podría ser computacionalmente costoso, y por ende, ser un proceso sumamente tardado con las herramientas disponibles. Por otra parte, si se tomaron en cuenta algunos aspectos de estos métodos para hacer la mejor predicción posible, tal es el caso del uso de validación cruzada sobre una cuadrícula de parámetros por modelo, con la finalidad de seleccionar aquellos que optimizarían el rendimiento de éstos.

Por último, y como se mencionó en el capítulo 5, no hay una técnica de clasificación que sea absolutamente mejor que otra para todos los conjuntos de información. Además habría que tomar en cuenta los recursos computacionales disponibles para poder llevar a cabo este trabajo a la práctica. De manera personal, los modelos que llevaráa a un ambiente de producción serían tanto las regresiones regularizadas, como las técnicas de bosques aleatorios y las máquinas de soporte vectorial. Sin embargo, habría que hacer algunos ajustes para poder aligerar el entrenamiento de las máquinas, ya que resultaría poco práctico utilizarlas con sus características actuales, donde el resultado de una sola corrida de validación cruzada tarde 3 días, y en producción, la predicción tendría que estar lista a lo más en horas, para ser oportuna.

Bibliografía

Ballings M., Van den Poel D., Hespeels N., Gryp R. (2015). *Evaluating multiple classifiers for stock price direction prediction.* Elsevier.

Berk R.A. (2017). *Statistical Learning from a Regression Perspective.* Springer, 291-308.

Chapman & Hall (2015). *Data Classification Algorithms and Applications.* CRC Press, 73-75, 67-70, 194-196.

Conover W.J. (1980). *Practical Nonparametric Statistics.* Wiley, 127-130.

Neter J. (1990). *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs.* CRC Press, 9-27.

Gibbons J.D. (2003). *Nonparametric Statistical Inference.* Marcel Dekker, 268-278, 422-428.

Kara Y. (2011). *Predicting directions of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange.* Elsevier.

Krstajic D., Ljubomir J.B., Leahy D.E., Thomas S. (2014). *Cross-validation pitfalls when selecting and assessing regression and classification models* Journal of Cheminformatics.

Hollander M., Wolfe D.A. (1973). *Nonparametric Statistical Methods.* John Wiley & sons, 50-54, 249-250.

Matloff N. (2017). *Statistical Regression and Classification from Linear Models to Machine Learning.* CRC Press, 65-96.

Montgomery D.C. (2006). *Introduction to linear regression analysis.* John Wiley & sons, 12-53, 67-119.

Murphy K.P. (2012). *Machine Learning A Probabilistic Perspective.* The MIT Press, 3-8, 9-13, 225-228, 245-252, 488-493, 496-505.

Murty M.N. (2016). *Support Vector Machines and Perceptrons Learning, Optimization, Classification and Applications to Social Network.* Springer, 7, 41-54.

Patel J. (2014). *Predicting stock and stock prive index movement using Trend Deterministic Data Preparation and machine learning techniques.* Elsevier.

Patel J. (2014). *Predicting stock and market index using fusion of machine learning techniques.* Elsevier.

Pathak M.A. (2014). *Beginning Data Science with R.* Springer, 87-99, 117-129, 131-132.

Perlin M.S. (2017). *Processing and Analysing Finacial Data with R.* Agencia Brasileira de ISBN, 193-203.

Rocjard J.R. (2020). *Just Enough R! An Interactive Approach to Machine Learning and Analytics.* CRC Press, 79-85, 99-108, 109-114, 280-283.

Siegel S., Castellan N.J. (1973). *Nonparametric Statistics for the Behavioral Science.* McGraw Hill, 75-82, 175-178.

Skiena S.S. (2017). *The Data Science Design Manual.* Springer, 213-218, 286-287, 354-356.

Tibshirani R., James G., Witten D., Hastier T. (2017). *An introduction to Statistical Learning with Applications in R.* Springer, 130-137, 163-164, 190-196, 303-330, 338-367.

Black Rock (08-11-2020). *Explicación de los ETFs.* `https://www.blackrock.com/mx/intermediarios/educacion/etf/explicacion-de-los-etfs`

Bolsa Mexicana de Valores (11-02-2021). *El Índice de Precios y Cotizaciones y su importancia para el mercado.* `https://blog.bmv.com.mx/2019/03/el-indice-de-precios-y-cotizaciones/`

Corporate Finance Institute (11-02-2021). *Technical Analysis A Begginers's Guide.* `https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/technical-analysis/`

Economipedia (08-11-2020). *Hueco de Ruptura.* `https://economipedia.com/definiciones/hueco-de-ruptura.html`

Economipedia (08-11-2020). *Indicador Momentum.* `https://economipedia.com/definiciones/indicador-momentum.html`

Economipedia (08-11-2020). *Media Móvil.* `https://economipedia.com/definiciones/media-movil.html`

Economipedia (08-11-2020). *Resistencia.* `https://economipedia.com/definiciones/resistencia.html`

Economipedia (08-11-2020). *Soporte.* https://economipedia.com/definiciones/soporte.html

María Sanchez (01-08-2020). *El S&P/BMV IPC cumple 40 años.* https://www.spglobal.com/spdji/es/documents/research/research-the-sp-bmv-ipc-turns-40-spa.pdf

Monex (11-02-2021). *Lo que debes saber del Índice INMEX.* https://blog.monex.com.mx/lo-debes-saber-del-indice-inmex

Rankia (11-02-2021). *¿Qué es el INMEX?.* https://www.rankia.mx/acciones/indice-precios-cotizaciones-ipc/blog/3127662-que-inmex

StockCharts (08-11-2020). *Stochastic Oscillator.* https://school.stockcharts.com/doku.php?id=technical_indicators:stochastic_oscillator_fast_slow_and_full

Visual Capitalist, Jeff Desjardins (02-05-2017). *12 Types of Technical Indicators Used by Stock Traders.* https://www.visualcapitalist.com/12-types-technical-indicators-stocks/

Wikipedia (11-02-2021). *S&P 500.* https://en.wikipedia.org/wiki/S%26P_500_Index

Wikipedia (11-02-2021). *S&P Global Ratings.* https://en.wikipedia.org/wiki/S%26P_Global_Ratings

Wikipedia (11-02-2021). *Stock market index.* https://en.wikipedia.org/wiki/Stock_market_index

Wikipedia (11-02-2021.) *The Vanguard Group.* https://en.wikipedia.org/wiki/The_Vanguard_Group